Metacomprehension Accuracy of Health-Related Information

Sarah Jayne Chadwick

Department of Psychology

Lancaster University

This thesis is submitted for the degree of

Doctor of Philosophy

December 2023

Abstract

As part of the production of written information, patient reader panels provide judgments of their understanding to evaluate the comprehensibility of draft documents. Previous research has suggested i) that there is a limited association, on average, between judgments of understanding and the comprehension demonstrated in tests of understanding and ii) that there is considerable variability between individuals in the direction and magnitude of this association. Unfortunately, while previous research implies, critically, that reader judgments of comprehensibility have limited utility, this research itself is characterized by important limitations that prevent firm conclusions.

This thesis comprises three experimental studies. The study design, method of measurement, and the approach to analysis were motivated by a critical review of previous research. The specification of participant, text and question sample sizes was determined by a novel method of prospective study design analysis, evaluating the accuracy and precision in effect estimation. The robustness of effect estimates are established through the series of empirical replications and in analytical sensitivity checks.

Across the studies, a weakly positive association between perceived and assessed comprehension was found across individuals, on average. Differences in reading ability and background knowledge did not reliably influence metacomprehension accuracy. Further, metacomprehension judgements were similarly predictive of performance on comprehension questions that targeted more versus less semantically central information. In contrast, metacomprehension judgements targeting specific ideas within texts were more predictive of understanding.

The findings of this thesis indicate that metacomprehension judgements are not a gold-standard method of evaluation: judgements show some predictive validity of comprehension outcomes, yet provide little insight into whether critical elements of the

documents are sufficiently understood. Overall, whilst situated within an applied context, the present research contributes more widely to the metacomprehension literature, making clear the need for a shift from traditional analytical approaches, in addition to greater theoretical precision.

Contents

List of Tables

List of Figures

Page:

Acknowledgements

I consider myself incredibly fortunate to have benefited from the guidance, encouragement and support of my two supervisors (Dr Rob Davies and Dr Debbie Costain). I extend my thanks to both for their expertise and enthusiasm which have motivated me to continuously improve my own understanding. The research reported herein has benefited substantially from their input.

I am also grateful for the support of my family and friends, and more broadly to all those who have helped me along this path. I also express my gratitude to the Economic and Social Research Council for the scholarship awarded in support of this research.

Declaration

I declare that the work presented in this thesis is, to the best of my knowledge and belief, original and my own work. The material has not been submitted, either in whole or in part, for a degree at this, or any other university. Approval for the submission of the thesis in excess of the standard length of 80,000 words was granted by the Pro-Vice-Chancellor (Education), acting on behalf of Senate, up to an additional 10,000 words. The approximate word count, excluding the material preceding and following the main text of the thesis is 88,653.

Name:	Sarah Jayne Chadwick
Date:	10/12/2023

1. Introduction

As part of our healthcare, we regularly receive information about health in written form. From cancer screening mailshots to outpatient care instructions, understanding these texts increases patients' capacity to make informed decisions, in addition to promoting engagement and encouraging positive health outcomes (Audit Commission, 1993; Department of Health, 2003). It is unsurprising, therefore, that such documents are carefully constructed according to specified guidelines (Department of Health, 2003) and undergo evaluation via reader panels prior to publication (e.g., Cambridge University Hospitals NHS Foundation Trust, 2023). Reader panels consist of individuals who volunteer their time to review written health-related information. Although practices vary by healthcare authority, reader panels typically consider how well the documents can be understood. Reader panellists, therefore, make judgments concerning their perceptions of text comprehension in order to evaluate the utility of written health information. Yet, in practice, our perceptions of text comprehension may not closely reflect the reality of our understanding.

Research suggests that our ability to discriminate between information we understand well, and that which we do not, is surprisingly poor (Prinz et al., 2020a; Yang et al., 2022), raising questions concerning the validity of reader panellists' judgements. Numerous studies have suggested that the alignment between perception and performance measures of comprehension can be influenced by various factors (Lin & Zabrucky, 1998; Prinz et al., 2020a). To date, however, the alignment between actual and perceived comprehension, in the context of health-related information, has not been directly examined. Addressing this gap in our knowledge, the research reported in this thesis examines the accuracy of judgements of comprehension made in respect of health-related texts and, in doing so, considers whether these judgements are likely to be useful in creating understandable patient information. In the remainder of this chapter, the research topic is first located within the broader context of metacognition and comprehension monitoring literature, before discussing theoretical perspectives and previous empirical work.

1.1 Metacognitive Judgements of Comprehension

Reflecting on our own comprehension is fundamentally a metacognitive process (Livingston, 2003). Metacognitive processes take any feature of cognitive activity as inputs (Flavell, 1976, 1978). Flavell (1979) distinguishes two features of metacognition which contribute to monitoring the progress of goal-oriented cognitive processes: knowledge, referring to our declarative knowledge or beliefs about our cognition; and experiences, referring to internal experiences generated before, during or after engaging in an activity. Metacognitive knowledge shapes our expectations of conditions which promote success, while metacognitive experiences provide subjective feedback about the progress or likely progress of an activity towards achieving a goal (Flavell, 1979). Metacognition can therefore be defined as a framework of processes which operate to enhance the efficacy of goal-directed cognitive activities through monitoring and control (Flavell, 1976, 1978, 1979). As an extension to this framework, 'metacognitive monitoring' is often used to refer jointly to evaluative and regulatory processes which modulate cognitive activities (e.g., Nelson & Narens, 1990; Nelson et al., 1994), though these processes may also be viewed as separate facets of metacognition, referred to as metacognitive monitoring and metacognitive control, respectively (Dunlosky & Metcalfe, 2009; Serra & Metcalfe, 2009).

Considering metacognitive processes in the context of text comprehension, 'metacomprehension' can be used to refer to the knowledge, evaluation and control processes related to comprehension (Brown, 1977; Lin & Zabrucky, 1998). Metacognitive knowledge concerning comprehension, or metacomprehension knowledge, can be further categorised into different types of knowledge, including person, task and strategy-related metacognitive knowledge (Flavell & Wellman, 1977; Flavell, 1979). Person-related metacomprehension knowledge refers to beliefs concerning oneself and others, such as believing that you read most effectively in quiet environments (Livingston, 2003). Task-related metacomprehension knowledge refers to task and processing demand beliefs, for example, recognising that it may take you less time to read familiar material (Myers & Paris, 1978). Lastly, strategy-related metacomprehension knowledge refers to knowledge of available strategies which may help to achieve the goal of a cognitive activity, such as rereading text to resolve comprehension difficulties (Lin & Zabrucky, 1998).

Metacognitive evaluation of comprehension is variously referred to as 'metacognitive monitoring of comprehension', 'metacomprehension monitoring', or simply 'comprehension monitoring', though these labels are used to also refer to metacognitive control processes (Baker, 1979, 1985; Baker et al., 2014; Griffin, Wiley & Thiede, 2019; Hacker, 1998; Keener & Hacker, 2012). For clarity, the term 'comprehension monitoring' will be used to refer to the metacognitive evaluation processes relating to comprehension, which generate experiences such as confusion when encountering comprehension difficulties (Flavell, 1979). In contrast, metacognitive control, or regulation, of comprehension will be used to refer to processes which are triggered when comprehension is found to be unsatisfactory following evaluation, such as re-reading or continuing reading to seek clarification (Bereiter & Bird, 1985). Regulation behaviours are themselves triggered by the evaluative processes of comprehension monitoring (Baker, 1985; Hacker, 1998; Zabrucky et al., 1993). The present research is primarily concerned with comprehension monitoring, the metacognitive evaluation processes associated with comprehension, rather than with regulatory processes.

1.2 Comprehension Monitoring

Comprehension monitoring can be understood as operating at two levels, with evaluations pertaining to either the processes involved in or to the products of text comprehension. During comprehension-building processes, comprehension monitoring occurs continuously to verify the coherence of the developing network of connections between extracted meanings and knowledge (Hacker, 1998; Grabe et al., 1987; Kendeou & van den Broek, 2005). For example, experiencing a disruption to your comprehension upon encountering an unfamiliar word is a metacognitive experience arising from comprehension monitoring processes which occur during reading. Comprehension monitoring which is oriented toward evaluating the understanding and learning gained from the episode of reading can also occur after reading or when reading is paused (Flavell, 1979; Glenberg & Epstein, 1985). For example, reflecting that reading a source of information did not sufficiently answer a question you sought to address is a metacognitive experience generated by comprehension monitoring after reading. Both forms of comprehension monitoring correspond to evaluation oriented toward achieving different but associated goals, capable of initiating metacognitive control mechanisms to regulate comprehension.

How readers keep track of their understanding of text is not a new topic of study (Baker & Brown, 1980). In the early twentieth century, Thorndike (1917) observed that children may appear to read fluently and engage satisfactorily in comprehension-building processes, yet incorrectly respond to comprehension questions. However, it was not until the late 1970s that research began to emerge which directly considered the accuracy of comprehension monitoring. Markman's (1977, 1979) research is cited as the seminal work exploring comprehension monitoring in developing readers, credited with establishing a methodology referred to as the 'error detection paradigm'. This methodology involves reading or listening to a text containing an anomaly, where observations of the capacity to detect the anomaly provide a measure of comprehension monitoring ability (Baker, 1989, 2014; Winograd & Johnston, 1980). This approach investigates the accuracy of comprehension monitoring which occurs during the comprehension-building processes of reading.

In Markman's (1977) first set of studies, children aged between six and nine years old were asked to help the experimenter check a set of instructions for a game and a magic trick. Key omissions meant that the instructions were incomprehensible and could not be followed. Younger children were found to require more probe questions before they reported a problem with the instructions, if at all (Markman, 1977). In a second set of studies, Markman (1979) examined the ability to report contradictions in factual texts about animals in children aged between eight and twelve. Children's comprehension monitoring was found to be poor, with approximately half of the children reporting a quarter, or less, of all errors (Markman, 1979). Based on these findings, it was suggested that failing to identify gaps in understanding was driven by an insufficient depth of processing, with children adopting lower standards for evaluating comprehension unrelated to the coherence of multiple pieces of information (Markman, 1977; 1979).

As poor comprehension monitoring had been demonstrated in developing readers only, early research also examined comprehension monitoring in older students. Baker (1979) hypothesised that comprehension monitoring may be poor even amongst skilled readers because monitoring may require knowledge and expertise which is not explicitly taught (Brown & DeLoache, 1978). In an error detection task, undergraduate students reported noticing only 23% of inconsistencies after reading and, despite then being instructed to search for them, were subsequently able to identify only 38% of the inconsistencies which went unnoticed during initial reading (Baker, 1979). Concerningly, there was little variation in the ratings of the comprehensibility in the task, with coherent texts and texts with inconsistencies perceived to be equally understandable (Baker, 1979). Consistent with Markman (1977, 1979), Baker (1979) concluded that the low levels of detection occurred due to students failing to spontaneously engage in sufficient comprehension monitoring during reading. Based on these results, comprehension monitoring problems appear not to be limited to children. Skilled readers apparently also fail to engage voluntarily in comprehension monitoring at an appropriate level during reading (Baker, 1979).

Consistent with the findings of Markman (1977, 1979) and Baker (1979), further research which has employed the error-detection paradigm has broadly indicated that the ability to detect and report inconsistencies amongst adult and developing readers can be poor (Baker, 1984; 1985; Baker & Anderson, 1982; Canney & Winograd, 1979; Epstein et al., 1984; Garner, 1980; Glenberg et al., 1982; Grabe et al., 1987; Paris & Myers, 1981; Winograd & Johnston, 1980; Zabrucky et al., 1987). Such ineffective comprehension monitoring during reading can lead readers to perceive incoherent texts as comprehensible, a phenomenon Glenberg et al. (1982) termed an 'illusion of knowing'. Not only do readers appear to regularly fail to spontaneously monitor their comprehension during reading, this lack of evaluative diligence generates inaccurate perceptions of text comprehensibility, thus undermining the veracity of their subsequent judgements of text consistency. In the context of the present research, concerning the accuracy of perceived comprehension judgements made of health-related text, therefore, this suggests that such judgements may produce misleading endorsements of text comprehensibility.

It would be premature, however, to conclude that readers are predominantly unable to accurately evaluate their comprehension. It has been argued that the error detection paradigm suffers major methodological flaws which render the results of studies employing it highly problematic in their interpretation. In the earliest critique of the paradigm, expanding on shortcomings discussed by Markman (1979) and Baker (1979), Winograd and Johnston (1980) identified five major criticisms of the methodology. The central issue is that a failure to report an inconsistency after reading does not equate to a failure to engage in comprehension monitoring during reading. There are many reasons why an inconsistency is not reported after reading, for example, the reader may somehow resolve the inconsistency during reading (Baker, 1979; Garner, 1987; Hacker, 1998; Markman, 1979; Winograd and Johnston, 1980). In effect, comprehension monitoring may successfully occur during reading, triggering regulatory processes which ensure the construction of a coherent understanding of the text. Judgements concerning the comprehensibility of a text, therefore, correspond to both the evaluative and regulatory aspects of metacomprehension. If a reader is able to make sense of a text, despite its inconsistencies, the text would be judged as comprehensible. To overcome these issues, Winograd and Johnston (1980) suggested using methodologies which provide a clearer link between processes and performance measures, explicitly considering the reader's purpose for reading and measuring comprehension monitoring during the reading process.

Concerns surrounding the use of the error detection paradigm gave rise to an alternative methodological approach, strongly influenced by metamemory research (Arbuckle & Cuddy, 1969; Wiley et al., 2005). In a paradigm which has sometimes been referred to as 'calibration of comprehension' (Lin & Zabrucky, 1998; Pilegard & Meyer, 2015), readers make judgements concerning their perceived understanding gained from coherent texts and, subsequently, answer comprehension questions. The correspondence between perceptions of comprehension and performance measures of comprehension, often referred to as 'metacomprehension accuracy' (Maki, 1998; Griffin, Mielicki & Wiley, 2019) or also as 'calibration of comprehension' (Glenberg & Epstein, 1985; Lin & Zabrucky, 1998; Weaver, 1990), provides an index of comprehension monitoring ability. This methodological approach investigates the accuracy of comprehension monitoring processes which operate on the products of the reading process. For clarity, the term 'metacomprehension task' will be used to refer to the methodology of using coherent texts to obtain perception and performance measures of comprehension, and the term 'metacomprehension accuracy' will be used to refer to the correspondence between these measures.

1.3 Metacomprehension Accuracy

Since the inception of the standard metacomprehension task paradigm, the approach taken to quantify the correspondence between measures of perception and performance has varied considerably. The measurement of metacomprehension accuracy will be considered in depth in Chapter 2 but is discussed here to aid in interpreting previous research. Briefly, there are two main approaches to measure metacomprehension accuracy, termed 'absolute accuracy' and 'relative accuracy' (Dunlosky & Lipko, 2007; Griffin, Mielicki & Wiley, 2019). Absolute accuracy captures the proximity between the magnitude of perception and performance scores, providing a measure of the precision of judgements (Bol & Hacker, 2012). In contrast, relative accuracy measures the systematic variation between perception and performance, capturing an individual's ability to discriminate between levels of understanding across multiple texts (Nelson, 1996). Within metacomprehension research, absolute accuracy is often measured as either the total signed or unsigned difference between perception and performance measures (Bol & Hacker, 2012), while relative accuracy is most frequently measured using intra-individual gamma correlations *G*, an ordinal measure of association which expresses the probability of directional correspondence between pairs of datapoints (Nelson, 1984). It has been suggested that absolute and relative accuracy capture different aspects of metacomprehension and are conceptually and statistically distinct (Bol & Hacker, 2012; Dunlosky & Lipko, 2007; Griffin, Mielicki & Wiley, 2019; Nelson, 1996). Researchers have argued that measures of relative accuracy are more appropriate (Griffin, Mielicki & Wiley, 2019; Nelson, 1984), while others have proposed reporting both absolute and relative measures (Schraw, 2009a).

The earliest research which employed the metacomprehension task, incorporating both absolute and relative measures of metacomprehension accuracy, suggested that readers appear to also be poor at evaluating the products of the comprehension-building processes.

Examining the accuracy of participants' predictions of performance, Maki and Berry (1984) asked college students to judge the probability of correctly answering multiple-choice questions after reading textbook chapters. Estimated absolute accuracy demonstrated overconfidence in the magnitude of performance predictions. The average magnitude of judgements concerning texts for which the comprehension questions were subsequently answered incorrectly were typically no lower than 'maybe correct'. Pearson correlations (*r*) between question performance and ratings were calculated, according to a median-split in performance, to examine relative accuracy. For those scoring above the median on the comprehension questions $r = .15$, whilst for those below the median $r = -.03$. Similar correlation estimates were observed in a second experiment (Maki & Berry, 1984). The findings suggested that, although higher performing readers may show greater discrimination in their judgements, individuals are generally poorly calibrated in both relative and absolute measures of metacomprehension accuracy.

At around the same time, Glenberg and Epstein (1985) likewise demonstrated readers' apparent inability to accurately evaluate comprehension after reading. In a series of three experiments, participants read single-paragraph texts, on a range of factual topics, and judged their ability to generate inferences based on the texts. Participants were then required to classify experimenter-generated inferences as true or false. Calibration curves, created using the aggregated proportion correct for each confidence rating, indicated both underconfidence in low ratings and overconfidence in high ratings. The highest rating of confidence was typically associated with proportions correct of less than 0.85. The lowest rating of confidence, however, was associated with proportions correct greater than 0.5. In addition, the average participant-wise point-biserial correlation r_{pb} between confidence ratings and response accuracy ranged from $r_{\text{pb}} = .04$ to .12 (r_{pb} is a measure similar to a Pearson correlation, but in which one variable is dichotomous). Considerable variability in these average correlations was observed between participants, with approximately 44% of participants estimated to have a negative correlation, whereas $r_{\text{pb}} > .6$ for some participants.

Given the large proportion of individuals who were either over-confident, or whose confidence ratings were apparently independent of their comprehension, it is understandable that the preliminary work by Maki and Berry (1984) and Glenberg and Epstein (1985) generated much interest in metacomprehension accuracy. Numerous studies have since reported that the accuracy of reader's judgements of their comprehension, particularly in terms of relative accuracy, is broadly poor (Dunlosky & Lipko, 2007; Griffin et al., 2009; Keener & Hacker, 2012; Lin & Zabrucky, 1998; Pilegard & Mayer, 2015; Thiede, et al., 2009; Wiley et al., 2016). In addition to the findings of the individual studies, three observations are often cited as evidence in support of poor metacomprehension accuracy: Maki (1998) found that the average gamma correlation across 25 studies was $G = .27$; Dunlosky and Lipko (2007) reported an average gamma correlation of $G = .27$ across 36 experimental conditions; and the findings of research reviewed by Lin and Zabrucky (1998) are consistent with an average correlation magnitude of approximately $G = .2$. More recently, comprehensive meta-analyses have estimated the average correlation magnitude to be $r = .24$ (Prinz et al., 2020a) and $G = .18$ (Yang et al., 2022), indicating further support for poor metacomprehension accuracy amongst skilled readers.

1.3.1 Theoretical Perspectives on Metacomprehension Judgements

The converging empirical evidence has prompted researchers to explore underlying causes of the typically poor metacomprehension accuracy observed across the population. Prevailing theoretical accounts of metacomprehension judgements claim that individuals rely on cues to make inferences about the quality of their understanding and poor accuracy results from the selection of cues which are not valid predictors of text comprehension (Griffin, Mielicki & Wiley, 2019; Koriat, 1997; Rawson et al., 2000; Thiede et al., 2010). Accounts proposing an inferential route to judgement partly emerged from Koriat's (1997) cue utilisation framework, which concerns judgements of learning for word pairs. Koriat (1997) describes three types of cues, intrinsic, extrinsic, and mnemonic, which overlap with Flavell's (1979) descriptions of metacognitive knowledge and metacognitive experiences (Griffin et al., 2013). Intrinsic and extrinsic cues are based on beliefs about how various aspects relate to learning, with the former corresponding to attributes of stimuli while the latter relates to features of the episode of learning (Koriat, 1997). For example, in the context of learning from text, the length of the text and the noise levels in the learning environment would be examples of intrinsic and extrinsic cues, respectively. In contrast, mnemonic cues relate to experiences generated during or after learning, such as the ease of processing or accessibility of information in memory. In forming judgements, a mix of these cues may be drawn upon and the selection of cues may vary between individuals and the situation (Koriat, 1997).

Building on Koriat's (1997) work and drawing on theories of text comprehension (Graesser et al., 1997; Kintsch, 1988), Dunlosky, Rawson and Hacker (2002) proposed the levels-of-disruption hypothesis to account for metacomprehension judgements. According to the levels-of-disruption hypothesis, disruptions during comprehension-building processes, or 'processing failures' (Rawson et al., 2000), such as encountering an unfamiliar word, are a major cue in forming metacomprehension judgements. The levels-of-disruption hypothesis makes three assumptions concerning how disruptions relate to metacomprehension judgements. Consistent with Koriat (1997), the inference and accuracy assumptions state that metacomprehension judgements are cue-based and that metacomprehension accuracy is a function of the correspondence between cues and test performance. The third assumption, the representation assumption, states that disruptions to processing at the lexical level or text base level will be less informative of comprehension than those at the situation model level, which constitutes an interconnected network of meanings, or ‘situation model’, drawn from the text (Kintsch, 1988; van den Broek et al., 2013; van Dijk & Kintsch, 1983).

Poor metacomprehension accuracy, according to the levels-of-disruption hypothesis, occurs due to a failure to process the text at the level which corresponds to that assessed in the comprehension questions (Dunlosky, Rawson & Hacker, 2002). As reading may be variously aimed at different levels of text comprehension (van den Broek et al., 2013; van Dijk & Kintsch, 1983), for readers who process the text at lower levels of representation, such as the lexical or text base level, disruptions to comprehension at the situation model level will go unnoticed (Dunlosky, Rawson & Hacker, 2002). As a result, for these individuals, cues which are highly predictive of performance on the comprehension task will be unavailable. In support of this account, Helder et al. (2016) found an inconsistency effect on reading times which preceded correct judgements of text coherence, but this effect was not observed where participants incorrectly judged the text to make sense. For readers who failed to detect a coherence-break during reading, this information was not available to inform the consistency judgement, leading to perceived text consistency (Helder et al., 2016). More generally, consistent with the levels-of-disruption hypothesis, research utilising the error detection paradigm has shown that behaviours during reading, including sentence reading times, eye-movement and neural activation, are associated with reporting problems with text coherence after reading (Ferstl et al., 2005; Helder et al., 2016; Helder et al., 2017).

Despite it’s potential explanatory utility, the levels-of-disruption hypothesis is limited in its capacity to fully account for metacomprehension judgements. Dunlosky, Rawson and Hacker (2002) highlight the lack of sufficient detail concerning how multiple sources of information may be weighted or integrated in the formation of metacomprehension judgements. In addition, the levels-of-disruption hypothesis indicates that disruptions to any of the levels of processing are identical and separable, yet this may not be the case (Kintsch, 1988). Encountering an unfamiliar word, for example, may be less disruptive to the coherence of the situation model than referential ambiguity, or vice versa, depending on the context within the text. Further, while a disruption may be detected, it may be resolved during the reading process. For example, breaks in text coherence may be repaired by the reader through drawing elaborative inferences (Baker, 1979; Winograd & Johnston, 1980). In these situations, therefore, experiencing a disruption during text comprehension is not a valid predictor of impaired understanding at the situational model level.

An alternative conception of metacomprehension judgements, influenced by the application of decision-making research (Tversky & Kahneman, 1974) to judgements of learning for word pairs (Scheck & Nelson, 2005), has been proposed by Zhao and Linderholm, (2008). The anchoring and adjustment model of metacomprehension accuracy posits that judgments result from a cue-based process of anchoring and adjustment. Judgement anchors are based on existing knowledge and may be spontaneously self-generated, such as perceived ability, or selected according to environmental features, such as experimental instructions (Epley & Gilovich, 2004; Linderholm et al., 2008; Zhao & Linderholm, 2008). Adjustments to these anchors are made using experience-based cues, such as reading ease or experience with the performance task (Kubik et al., 2022; Linderholm et al., 2008; Zhao, 2022). According to this model, the use of knowledge-based cues as anchors is driven by uncertainty, such as a lack of information about the performance task. The extent to which anchors are adjusted is dependent on an individual's motivation to engage in effortful adjustment processes (Epley & Gilovich, 2006; Zhao & Linderholm, 2008). Importantly, as some experience-based cues lack predictive validity, such as processing ease, adjustments processes are not reliably beneficial (Zhao & Linderholm, 2008). Either insufficient or inaccurate revisions to knowledge-based anchors result in judgements which do not correspond to performance on the performance task, reducing metacomprehension accuracy (Zhao & Linderholm, 2008).

While Zhao and Linderholm's (2008) anchoring and adjustment model of metacomprehension accuracy can successfully explain why judgements can show stability across texts (Moore et al., 2005), this account does not provide a clear description of either the anchoring or the adjustment processes. While various cues which may serve as anchors are proposed, the process through which individuals select from competing anchors is not described. Similarly, while cues which may be used to adjust anchors are identified, no description is provided regarding how this information is combined with anchors. Given this lack of specificity, in considering observed changes to judgement magnitudes in response to an experimental manipulation, it is unclear whether such changes can be attributed to a difference in anchor selection or reflect the use of adjustment processes. More generally, the distinction between anchors as knowledge-based and adjustments as experience-based is arguably contextually dependent, as the latter may also readily be used as a judgement anchor (Epley & Gilovich, 2004).

More recently, alternative accounts of metacomprehension judgements have been proposed which also draw on both Koriat's (1997) cue utilisation framework and Kintsch's (1998) construction-integration model of text comprehension. Griffin et al. (2009) proposed a dual-route model, specifying that individuals use either heuristic or text representation cues to form metacomprehension judgements. Similar to Koriat's (1997) conceptualisation of cues, heuristic cues are available regardless of whether a text is read, while text representation cues are generated through the experience of reading a text (Griffin et al., 2009; Thiede et al., 2009). Cues which pertain to the text representation can also be produced after reading using generative tasks, such as the recalling keywords or summarising the text (Griffin et al., 2009). As text representation cues are more closely related to the quality of the products of comprehension, these cues will be more predictive of understanding than heuristic cues (Griffin et al., 2009; Kintsch, 1998; Wiley et al., 2005). According to this account, therefore, poor metacomprehension accuracy results from the use of heuristic cues with low predictive validity (Griffin et al., 2009).

Expanding on Griffin et al.'s (2009) dual-route framework, Wiley et al., (2016) proposed what is now referred to as the 'situation-model approach to metacomprehension' (Griffin, Mielicki & Wiley, 2019). According to this account, the predictive validity of text representation cues vary depending on how closely these cues relate to the situation model level of understanding (Griffin, Mielicki & Wiley, 2019; Kintsch, 1998; Wiley et al., 2016). Only cues which relate to the integration of information from the text and background knowledge, in the construction of a situation model of the text, can accurately reflect understanding (Griffin, Mielicki & Wiley, 2019; Wiley et al., 2016). For example, reading a text with short words and sentences may generate an experience of reading fluency, but may not translate to high levels of comprehension. In contrast, a feeling of confusion when unsuccessfully summarising a text provides a metacognitive experience which is informative of comprehension. Observed poor metacomprehension accuracy, therefore, occurs when individuals fail to use cues which relate to the situation model level of comprehension (Wiley et al., 2016). In addition, consistent with Dunlosky, Rawson and Hacker's (2002) accuracy assumption, cues pertaining to the situation model will only be predictive of comprehension insofar as the comprehension test assesses this level of understanding (Wiley et al., 2016).

Notwithstanding the volume of evidence consistent with the beneficial effects of using situation-model level cues in forming judgements (Prinz et al., 2020b), both Griffin et al.'s (2009) dual route framework and Wiley et al.'s (2016) situation-model approach to metacomprehension are similarly limited in their capacity to fully explain the processes underpinning metacomprehension judgements. Wiley et al. (2016) implies that cue selection is knowledge driven, as readers rely on norms when forming judgements but can be trained to select appropriate cues. In contrast, Griffin et al. (2009) suggest that individuals typically select heuristic cues, as this may be less effortful and these cues may be more salient than text representation cues. Within both accounts, it is not clear whether or not readers make use of multiple cues when forming judgements. Although the use of multiple cues is not precluded, the processes through which integration or weighting of cues may occur is not described in detail. Further, if multiple cue-use occurs, it is unclear whether individuals may select from either heuristic or text representation cue types exclusively, as implied by Griffin et al., (2009), and, if so, why these cue types provide mutually exclusive routes to judgement (Koriat, 1997).

Overall, cue-based theories provide a more cogent account of metacomprehension judgements than accounts which suggest that individuals can directly evaluate the quality of internal stimuli states to form judgements (Arbuckle & Cuddy, 1969; Hart, 1967). Such direct access accounts fail to explain why judgements are only weakly predictive of performance, on average (Koriat, 1997; Schwartz, 1994; Tauber & Dunlosky, 2016). The theories discussed here describe both the causes of poor metacomprehension accuracy and the conditions under which accuracy can be improved. These theories are able to provide a coherent account of the observed individual variability in metacomprehension accuracy, ascribing this to differences in the availability, or use of, cues which are diagnostic of comprehension. However, as cue selection and integration processes are underspecified within these accounts, it is challenging to make firm predictions of a given individual's level of metacomprehension accuracy under particular judgement conditions. Since the predictiveness of metacomprehension judgements, and therefore the utility of reader panellists’ judgements, is fundamentally a concern which centres on an individual's metacomprehension accuracy, the key question then is: what are the factors which engender differences in metacomprehension accuracy?

1.3.2 Influences on Metacomprehension Accuracy

Emerging from the research conducted to date is evidence that metacomprehension accuracy may be sensitive to a range of variables, both external and internal to the individual (Prinz et al., 2020a). These variables can be categorised according to whether they relate to: i) the individual, such as differences in various skills and abilities, ii) the text, such as how texts are written or presented, iii) or the design of the metacomprehension task, including aspects of both methodology and measurement (Lin & Zabrucky, 1998; Prinz et al., 2020a; Schraw, 2009b; Weaver et al., 1995). In better understanding the conditions that affect metacomprehension accuracy, the remainder of this chapter will discuss empirical research which has investigated these variables. Given the volume of research conducted to date, the discussion which follows focuses on subsets of variables from each of these categories which are most pertinent to evaluating whether current evidence supports the utility of reader panel judgements.

Individual Differences. Despite evidence of considerable individual variability in metacomprehension accuracy, the factors which may be responsible for driving such differences have received comparatively less attention than task and text-related variables (Chiang et al., 2010; Zabrucky, 2010). In the context of reader-panels, understanding why a group of readers' metacomprehension accuracy can differ in response to identical texts and procedures is fundamental to identifying panelists who are equipped with the characteristics which engender accurate evaluations of patient information. While there is limited research examining the sources of individual differences in metacomprehension accuracy, two relevant variables which have previously been considered as potential moderators are reading ability and background (or subject) knowledge.

Reading Ability. Variation in reading ability has repeatedly been suggested to influence metacomprehension accuracy (Kurby et al., 2007; Lin et al., 2002; Maki et al., 2005; Ozuru et al., 2012). Given that monitoring comprehension during reading may be cognitively demanding (Chiang et al., 2010; Nelson & Naren, 1990), the resources available to readers who are less efficient in text processing may be limited. Lower ability readers may be unable to concomitantly attend to valid cues, concerning the quality of their comprehension, whilst reading (Griffin et al., 2008; Griffin et al., 2009), resulting in lower metacomprehension accuracy. In support of this claim, Griffin et al. (2008) found that reading ability and working memory capacity were significant predictors of metacomprehension accuracy. When these variables were included as predictors of participant-level Pearson correlations in a regression model, each independently explained variability in metacomprehension accuracy, indicating that reading ability and working memory capacity impose separate constraints on concurrent processing which limit metacomprehension accuracy. However, considerable variability in metacomprehension accuracy across the range of reading ability and working memory capacity scores was observed, indicating that scoring highly on these measures did not prevent inaccurate judgements nor did scoring poorly on these measures prevent accurate judgements.

Further support for the contention that lower ability readers may attend to cues which are not directly related to the reading experience can be observed in Kurby et al.'s (2007) research. In this study, participants completed the Gates-MacGinitie Reading Test (MacGinitie & MacGinitie 1989) to measure reading ability and provided sentence-by-sentence metacomprehension judgements on a text which was modified to vary the difficulty of sentences. The average magnitude of judgements for each sentence for low and high ability readers was regressed on sentence difficulty and average reading time, using a median split of reading ability score to determine this dichotomy. Sentence difficulty was found to predict judgements of learning for high ability readers, whereas judgements were unrelated to difficulty or reading time for low ability readers. Further, using a similar methodology, Ozuru et al. (2012), found that participant-level gamma correlations for sentence-level predictions of performance were, themselves, significantly correlated with reading ability (G = .53). In addition, regression analysis indicated that reading ability accounted for 28% of the variability in participant's gamma correlations (Ozuru et al., 2012).

In contrast to these findings, other research has failed to provide clear evidence of an effect of reading ability on metacomprehension accuracy. In research conducted by Maki et al. (2005), participants were grouped into three reading ability levels, based on verbal ability scores obtained from student records, and were asked to provided metacomprehension judgements on texts which were manipulated on lexical and grammatical complexity. Average gamma correlations were not observed to significantly differ between the levels of reading ability, regardless of text complexity. In addition, overconfidence in the magnitude of predictions of performance was observed across all reading ability levels on simplified texts. In addition, research conducted by Lin et al. (2002) demonstrated that age group did not significantly influence metacomprehension accuracy or interact with text readability, measured as variation in Flesch-Kincaid grade level (Flesch, 1948). Given that the average reading ability, measured using the Nelson-Denny reading test (Brown et al., 1981), was found to be higher amongst younger adults, Lin et al. (2002) surmised that readers with differing ability levels may be equally influenced by text difficulty.

An alternative perspective on the potential influence of reading ability on metacomprehension accuracy purports that it is not reading ability per se that affects accuracy, but self-perceptions of reading ability (Kwon & Linderholm, 2014). Accounting for the variability in metacomprehension accuracy across reading ability levels, Kwon and Linderholm (2014) proposed that self-perceived reading ability is a key component of metacomprehension judgements. Analysis of the discrepancy between the magnitudes of self-perceived reading skill and actual reading skill revealed that this difference significantly predicted the discrepancy between predictions and performance on the metacomprehension task. A greater distance between readers' beliefs and the reality of their reading skill, therefore, led to greater discrepancies between predicted and actual performance. Further analyses indicated that, on average, participants scoring below the 30th percentile over-estimated their reading ability, while participants above the 70th percentile underestimated their reading ability. While inaccurate self-perceptions of reading ability appear to reduce metacomprehension accuracy, these results suggest that this occurs across reading ability levels, rather than being limited to low ability readers. Given Kwon and Linderholm's (2014) and other's mixed findings (Griffin et al., 2008; Lin et al., 2002; Maki et al., 2005; Ozuru et al., 2012), the impact of reading ability on metacomprehension accuracy is unclear. However, discussion of the potential impact of reading ability will be revisited later in this chapter.

Background Knowledge. The reader's level of expertise in the topic of the text, or background knowledge, has also been suggested to influence metacomprehension accuracy. It has previously been claimed that high background knowledge permits greater accuracy in evaluating comprehension, through a better understanding of what it means to comprehend a text within one's field of expertise (Kruger & Dunning, 1999; Lin & Zabrucky, 1998; Weisberg et al., 2008; Wiley et al., 2005). However, early research by Glenberg and Epstein (1987) found that greater topic expertise may instead generate misleading feelings of confidence in understanding, relative to less familiar domains. Regressing participant-level gamma correlations on the number of topic-specific courses completed by participants, Glenberg and Epstein (1987) found that a higher number of completed topic-specific courses reduced metacomprehension accuracy on physics-related texts, while no effect was found on music-related texts. Overall, despite lower confidence, students often did not perform more poorly on topics they were less familiar with, suggesting that confidence ratings were based on expectations informed by familiarity (Glenberg & Epstein, 1987).

Research that has addressed a number of concerns with Glenberg and Epstein's (1987) study, such as the validity of the measure of expertise, has produced a somewhat similar pattern of results. In the context of baseball-related expertise, Jee et al. (2006) found that the correlation between participant-level Pearson correlations of text level predictions of performance and scores on a 45-item baseball knowledge questionnaire was nonsignificant (*r* = -.08). In contrast, the estimated relationship between expertise and absolute metacomprehension accuracy was positive (*r* = .25). Using the same stimuli and methodology, these findings were replicated by Griffin et al. (2009). Relative metacomprehension accuracy again showed a nonsignificant association with expertise (*r* = -.07), while absolute accuracy was positively correlated with expertise (*r* = .32). These findings indicate that greater topic knowledge is associated with lower average deviations between predictions and performance, yet does not help individuals reliably identify texts which they understand well and those which they do not.

More recent research has provided conflicting evidence regarding the impact of expertise on absolute metacomprehension accuracy, while replicating the nonsignificant effect of expertise on relative metacomprehension accuracy. Across two experiments, Shanks and Serra (2014) collected judgements of learning and measured recall performance for 100 single-sentence facts, drawn from 10 topics which participants ranked from best to least well-known. Average absolute accuracy was found to be lower for topics considered to be most well-known compared to less well-known topics, indicating that greater perceived topic knowledge increased overconfidence. Comparing gamma correlations between judgements of learning and question performance across the 10 topic rankings showed that average relative accuracy was similar across topics considered more or less well-known. Given these findings, and those discussed above, contrary to theoretical expectations, greater expertise seems not to improve individuals’ ability to provide metacomprehension judgements which accurately discriminate between texts understood more or less well. However, with respect to the difference in magnitude between predictions and performance, the impact of background knowledge appears less clear.

Text Characteristics. The second category of variables which have been suggested to influence metacomprehension accuracy relates to characteristics of the stimulus texts. Two aspects of the texts that have previously been considered are the difficulty of the text and the presence of features which are supplementary to the information in the text. In the context of reader panels, the degree of complexity in patient information and the presence of features, such as images or emboldened font, may influence the utility of the evaluations made by reader panelists.

Difficulty. Studies which have examined the effect of the difficulty of the text on metacomprehension accuracy have yielded somewhat mixed results but indicate that varying levels of text difficulty may present more of a challenge to accuracy for lower ability readers. In this research, difficulty is typically operationalised as the readability of the text, corresponding to differences in lexical and syntactical complexity or coherence (Prinz et al., 2020a). Alongside text difficulty, an individual's reading ability is often also considered due to the interdependence between these variables and comprehension. With regards to the effect of text difficulty on relative measures of absolute accuracy, Weaver and Bryant (1995) examined the effect of readability on metacomprehension accuracy using easy, intermediate and difficult texts, corresponding to Flesch-Kincaid grade levels (Flesch, 1948) of readability below 8th grade, around 12th grade and above 16th grade. Metacomprehension accuracy was found to vary as a function of text difficulty, with average gamma correlations of *G* = 0.29, *G* = 0.69 and *G* = 0.30 observed for easy, intermediate and difficult texts, respectively. Since participants were assumed to be at approximately the 12th grade reading level, Weaver and Bryant (1995) suggested that metacomprehension accuracy is greater when the difficulty of the text aligns with the reading level of the individual.

Other research has failed to replicate Weaver and Bryant's (1995) finding regarding the effect of text difficulty and reading skill on relative metacomprehension accuracy. Lin et al. (2002) found that, across reading ability levels, measured using a standardised reading test, metacomprehension accuracy was higher on moderate difficulty texts (*G* = .47) than on easy or hard texts (*G* = .19 and *G* = -.09). Readers with reading levels above the level of moderate texts were reported to show greater accuracy on these texts compared to hard texts, though it is not clear whether this difference was analysed statistically. In contrast, Maki et al. (2005) found no significant differences in gamma correlations between original and reduced-difficulty texts, regardless of participants' reading ability, with ability estimated using student record data. However, on revised texts, only the accuracy of high ability readers differed from zero (*G* = .48), suggesting that higher ability readers may be more able to accurately evaluate their understanding of lexically and grammatically simplified texts. More recently, Prinz et al. (2020a) found no evidence for a linear or inverted-U shape effect of variability in Flesch-Kincaid grade level on metacomprehension accuracy. Yet, as this finding was limited to college-level students, who may not show large variation in reading ability, the interaction between difficulty and reading ability was arguably not fully explored, (Prinz et al., 2020a).

Research which has considered the impact of text difficulty on absolute measures of metacomprehension accuracy has likewise provided inconsistent findings, yet similarly indicates that the accuracy of lower ability readers may be impacted by varying text difficulty. For example, Maki et al.'s (2005) findings suggest that overconfidence in judgements of comprehension is influenced by both text difficulty and reading ability. On reduced-difficulty texts, overconfidence was observed across reading ability levels, whereas only low ability readers showed overconfidence on texts with greater lexical and grammatical complexity. In contrast, Golke and Wittwer (2017) found that judgements on highly coherent texts, in which connections between information were explicit, were associated with underestimation of performance, while overestimation of performance was associated with less coherent texts. Greater reading skill appeared to decrease bias on less coherent texts, while reading skill did not influence bias on the highly coherent text (Golke & Wittwer, 2017). Despite the inconsistency regarding the impact of low difficulty texts, both Maki et al.'s (2005) and Golke and Wittwer's (2017) findings indicate that increased text difficulty reduces accuracy for lower ability readers.

Supplementary Features. Previous research has also explored how metacomprehension accuracy may be influenced by the presence of additional features within the text which can be considered supplementary to the written information. One such supplementary feature is the inclusion of illustrations. It has previously been suggested that including images in texts can increase perceptions of the quality and understandability of the information (Kools et al., 2006; McCabe & Castel, 2008; Serra & Dunlosky, 2010; Wiley, 2019), with informative diagrams able to promote learning compared to text alone (Butcher, 2006; Mayer, 2009). However, research suggests that the inclusion of illustrations may not influence relative metacomprehension accuracy. Serra and Dunlosky (2010) evaluated metacomprehension accuracy using texts containing either diagrams, photographs or no images. The magnitudes of average predictions of performance for both diagram and photograph conditions were found to be significantly greater than those for the text-only condition. However, only including a diagram produced significantly greater comprehension than either the text-only or the text featuring a photograph. Despite this pattern of results, there were no significant differences in the average gamma correlations between these conditions; the ability of readers to accurately discriminate their levels of understanding across texts was not influenced by the inclusion of photographs (*G* = .26) or diagrams (*G* = .12), compared to text alone (*G* = .16).

A second supplementary feature which has previously been considered is the inclusion of attention-signaling elements, with metacomprehension accuracy generally expected to improve where the reader's attention is directed to relevant information (Gier et al., 2009; Margolin, 2013). Consistent with these expectations Margolin (2013) demonstrated an increase in the magnitudes of average metacomprehension judgements and performance on texts in which relevant information was presented in a bold font. However, improvements have not been observed for relative metacomprehension accuracy. In Margolin's (2013) research, the average gamma correlations for texts including bold font did not significantly differ from zero, whereas the opposite was observed on texts without bold font. Similarly, Gier et al. (2009) demonstrated that metacomprehension accuracy was lower on texts in which information included on the comprehension test was highlighted (*G* = .27), compared to texts with no highlighting (*G* = .37). Gier et al. (2009) also found that highlighting information which was not later assessed was detrimental to metacomprehension accuracy (*G* = -.36). It would appear, therefore, that metacomprehension accuracy may be reduced by attention-signaling devices which do not provide additional information or emphasise irrelevant aspects of the text.

Metacomprehension Task Design. In seeking to identify the variables which may increase metacomprehension accuracy, methodological features that may alter the judgement process or increase the robustness of the measures of interest are frequently considered. These variables can be further categorised as associated with the format of the task, the perception measures of comprehension, or the performance measures of comprehension. The format of the task refers to features of how the calibration paradigm is deployed, perception measures of comprehension refer to the prompts used to elicit metacomprehension judgements, and performance measures of comprehension refer to the items used to assess understanding of the text. In the context of reader panels, each of these sub-classifications of variables reflect differences in how the document review process may be conducted and differences in how evaluations may be elicited that may influence the utility of judgements.

Task Format. A considerable volume of research has indicated that features of how the metacomprehension task is deployed can influence metacomprehension accuracy (Prinz et al., 2020a). One such variable which has been argued to be too important to be left to chance is the instructions given to participants before reading (Winograd & Johnston, 1980). The major variation in the instructions given to readers in the calibration paradigm is whether or not readers are informed that they will be taking a test on the material before reading or making metacomprehension judgements (Prinz et al., 2020a). Instructing readers to read for understanding (e.g., Lin et al., 2000, 2002; Weaver & Bryant, 1995) is somewhat less prevalent than informing participants of a test (e.g., Anderson & Thiede, 2008; Glenberg & Epstein, 1985; Griffin, Wiley & Thiede, 2019; Rawson & Dunlosky, 2002; Thiede et al., 2003; Wiley et al., 2008; Wiley et al., 2016), likely owing to the frequent use of predictions of performance as the measure of perceived comprehension..

Researchers have previously contended that a reader's purpose for reading can strongly influence the standards of coherence they apply during reading, impacting upon the types and strengths of connections made (van den Broek et al., 2011; van den Broek & Helder, 2017). Variation in instructions can, therefore, fundamentally shape an individual's conception of what it means to comprehend a given text (Wiley et al., 2005), affecting both the processes and products of comprehension. For example, when readers are instructed to read for study purposes, they report more coherence-building inferences and show greater memory during free recall, compared to reading for entertainment (van den Broek et al., 2001) Similarly, when instructed to read for the purpose of undertaking a task, readers' memory of the text is sensitive to how important the information is considered for completing the task (Mills et al., 1995). As a consequence, therefore, the alignment between perception and performance measures of comprehension may be altered on the basis of the instructions provided.

To date, research has yielded inconsistent findings regarding the impact on metacomprehension accuracy of informing participants of the requirement to perform an activity based on understanding the text. Early research by Schommer and Surber (1986) compared two instruction conditions: considering whether a text was written clearly or to teach the main points of the text. Inaccuracies in metacomprehension were measured by taking a count of instances in which participants scored below 50% on the test whilst judging their understanding at or above the midpoint of the metacomprehension rating scale. Reading to teach the main points of the text was found to increase metacomprehension accuracy, conditional on the text difficulty being greater than the reader's expected grade level. However, research conducted by Linderholm and Wilde (2010) found that instructions to read for the purpose of studying for an essay-based exam led to inflated perceptions of comprehension compared to reading for entertainment, for both judgements of comprehension and predictions of performance. More recently, the meta-analysis conducted by Prinz et al. (2020a) indicated that metacomprehension accuracy does not differ depending on whether readers are informed of an upcoming test of comprehension (r = .23) or not (r = .27).

The impact on metacomprehension accuracy of informing individuals of an upcoming test, however, may depend on the extent to which information about the nature of the assessment is provided. Griffin, Wiley and Thiede (2019) found that metacomprehension accuracy on inference-based comprehension questions is increased when participants are explicitly told that the test will assess the ability to connect pieces of information (*G* = .53), whereas this effect was not observed when participants were simply informed that a test would be administered (*G* = .16). In related research, prior to reading, Magliano et al. (1993) informed participants they would complete a comprehension test on each text after reading and instructed participants to either focus on making connections and summarising the text or to focus on word-sounds. It was found that participants instructed to use a superficial comprehension strategy showed poor metacomprehension accuracy during early texts but improved over subsequent texts as more information about the nature of the test became available (Magliano et al., 1993). Therefore, providing information about the type of test, not just the existence of a test, can improve metacomprehension accuracy.

In fact, a body of research indicates that metacomprehension accuracy can be improved by instructing individuals to engage in an intervention task which, similar to the nature of the assessment task, requires an explicit consideration of the knowledge and understanding gained from the text (Prinz et al., 2020b). For example, research by Griffin et al. (2008) found that instructing participants to explain the meaning of each sentence and how it relates to the text during reading was a significant, positive predictor of metacomprehension accuracy. However, for some interventions, the effect on metacomprehension accuracy appears to be conditional on a delay between reading and engaging in the intervention. Thiede et al. (2003) found that instructing participants to list keywords from texts after reading improved metacomprehension when there was a short delay (*G* = .70, approximately), compared to no intervention or immediately listing keywords (*G* = .35 and *G* = .40, approximately, respectively). Similarly, Thiede et al. (2010) demonstrated that writing a summary of the text following a delay improves metacomprehension accuracy (*G* > .6), compared to writing an immediate or no summary (*G*

< .3 for both conditions), for both normal readers and those requiring remedial reading classes.

Comprehension Perception Measures. The second type of task-related variables which may influence metacomprehension accuracy relates to the measurement of perceived comprehension. Across metacomprehension accuracy research, the wording of the prompts used to elicit comprehension judgements is highly variable, with studies frequently employing more than one type (Pilegard & Mayer, 2015; Schraw, 2009b; Thiede et al., 2009). For example, researchers may elicit judgements of learning (Pilegard & Mayer, 2015; Shanks & Serra, 2014), of understanding or comprehension (Linderholm & Wilde, 2010; Thiede & Anderson, 2003; Weaver & Bryant, 1995), of perceived ease or difficulty (Lin et al., 2001; Maki & Serra, 1992), or predictions of performance (Baker & Dunlosky, 2006; Glenberg & Epstein, 1985; Serra & Dunlosky, 2010; Weaver, 1990; Wiley et al., 2016). How these judgements are operationalised also varies. For example, judgements of understanding may be framed as a degree of confidence in understanding (Glenberg & Epstein, 1985; Maki et al., 1990) or as a quantification of the amount of text understood (Serra & Dunlosky, 2010). Research that has investigated how this variability may influence metacomprehension accuracy has so far provided mixed findings.

Research by Maki and Serra (1992) suggests that judgements which capture individuals' perceptions of their ability to perform on the comprehension task associate more strongly with assessed comprehension than judgements of the perceived ease of understanding. In two experiments, average gamma correlations based on predictions of performance were found to significantly differ from zero (experiment 2: G = .40; experiment 3: G = .25), whereas ease of comprehension judgements did not (experiment 2: G = .14; experiment 3: G = .11). In contrast, Ozuru et al. (2012) reported non-significant differences between the average gamma correlations of sentence-by-sentence judgements for ease of understanding (*G* = .09) and predictions of performance (*G* = .14). Lin et al. (2001) likewise observed highly similar average gamma correlations across judgements of understanding, confidence and ease: *G* = .15, *G* = .14 and *G* = .14, respectively. In addition, metacomprehension accuracy was not found to significantly differ between predictions of performance and judgements of understanding in a recent meta-analysis (Prinz et al., 2020a).

In a more fine-grained examination, Pilegard and Mayer (2015) explored the impact on metacomprehension accuracy of judgement type, the wording of response options and the nature of the comprehension questions. A non-significant difference in metacomprehension accuracy was found between judgements of understanding and judgements of learning. However, metacomprehension accuracy appeared greater for judgements of understanding on test items which required transferring information from the materials to explain new scenarios: *r* = .38 and *r* = .62 for judgements of learning and judgements of understanding, respectively. The main analysis of judgment wordings, including 'how much', 'how confident', 'how many' and 'how difficult', indicated no overall significant differences between the four wording types. Nevertheless, for both judgement types, on test items which required the transfer of information, metacomprehension accuracy did not exceed *r* = .30 for the response wording 'how difficult', on average, whereas for other wording types this was *r* > .40. Given these findings, Pilegard and Mayer (2015) suggested that judgements of understanding may be marginally more accurate than judgements of learning but that quantifying judgements in terms of difficulty may reduce accuracy. However, as the differences between these conditions did not achieve statistical significance, these claims are lacking in strong support.

A further source of variability in how measures of perceived comprehension are obtained relates to the amount of text the judgement corresponds to. Judgements made with respect to a whole text, or across multiple portions of text, are often referred to as 'global judgements' (Dunlosky & Lipko, 2007; Glenberg & Epstein, 1985; Griffin et al., 2009; Huff & Nietfeld, 2009; Linderholm & Wilde, 2010). In contrast, 'local judgements' refer to judgements concerning individual items (Nietfeld et al., 2005) such as a sentence, paragraph, concept or specific term (Dunlosky et al., 2005, Dunlosky & Lipko, 2007; Ozuru et al., 2012; Pilegard & Mayer, 2015; Serra & Dunlosky, 2010). It has been suggested that global judgements, which concern larger quantities of information than local judgements, induce comparatively lower metacomprehension accuracy (Dunlosky & Lipko 2007; Thiede et al., 2009). This may occur due to inaccuracy introduced in the process of making a single judgement concerning multiple items, for example, by taking an average of a sample of the information to be judged (Dunlosky et al., 2005; Händel & Dresel, 2018; Koriat, 1995; Lefèvre & Lories, 2004). Alternatively, it may be that judgements across multiple pieces of information are challenging to produce and, as a consequence, individuals may rely on cues which are not directly related to the reading experience (Dunlosky et al., 2005; Dunlosky & Lipko, 2007; Händel et al., 2020). Somewhat contrary to expectations, however, studies which have examined the impact of the judgement scope on metacomprehension accuracy have produced mixed findings.

Considering the relationship between judgement scope and relative metacomprehension accuracy, Dunlosky, Rawson and McDonald (2002) found a non-significant difference in the average gamma correlations for predictions of performance for term-specific and whole-text judgements (*G* = .42 and *G* = .40, respectively). However, in a follow-up study, Dunlosky et al. (2005) reported that the relative accuracy of participants' predictions of correctly recalling definitions of specific terms were more accurate than overall predictions of performance, conditional on engaging in a prejudgement recall attempt. In contrast, Lefèvre and Lories (2004) found that global judgements were more accurate than local judgements. The average gamma correlations for per-text judgements were found to be significantly greater than those made per-paragraph: G = 0.43 and G = 0.38 for text-level judgements, and G = 0.36 and G = 0.27 for paragraph-level judgements, in the first and second study, respectively). However, local judgements in Lefèvre and Lories' (2004) study related to multiple pieces of information, similar to global judgements, rendering the interpretation of the effect of judgement scope less clear.

Similarly, research which has examined the influence of judgement scope on absolute metacomprehension accuracy has provided inconsistent results. Schraw (1994) found that the absolute accuracy of post-diction judgements of performance, based on a single set of comprehension questions, did not differ from a single, overall judgement of performance across multiple sets of comprehension questions. Likewise, Dunlosky et al. (2005) demonstrated that average absolute accuracy did not differ between predictions of performance for defining specific terms, compared to predictions of performance concerning overall performance on a test of the text material. Contrary to these findings, Nietfeld et al. (2005) found that single judgements of confidence in performance across multiple test items were significantly more accurate than judgements of confidence in responding correctly to individual test items, suggesting that local judgements may increase the discrepancy between judgements and performance.

Comprehension Performance Measures. The third group of variables which relate to the metacomprehension task design concern the performance measures of comprehension. While an objective comprehension assessment is generally not a component of reader panel evaluations, the approach taken to measure comprehension in empirical research influences estimated metacomprehension accuracy and, therefore, the conclusions arising from such research. For instance, researchers have contended that metacomprehension accuracy can be affected by the number of comprehension questions used to measure understanding (Thiede et al., 2009; Weaver, 1990). In studies using a limited number of questions to assess comprehension, the resulting measure of understanding can lack reliability and validity, as readers are unable to sufficiently demonstrate their understanding gained from the text (Thiede et al., 2009; Weaver, 1990). Single-item measures of comprehension have been argued to be the most vulnerable to these issues, as comprehension is misleadingly constructed as an all-or-nothing state (Lin & Zabrucky, 1998; Weaver, 1990; Wiley et al., 2005). As a result, even well-calibrated individuals may appear to provide inaccurate judgements of comprehension (Weaver, 1990). Estimates of metacomprehension accuracy based on a limited set of questions, can, therefore, artificially constrain the alignment between perception and performance measures of comprehension (Dunlosky & Lipko, 2007; Weaver, 1990).

In support of these claims, research has demonstrated a positive relationship between metacomprehension accuracy and the number of comprehension questions. Examining the impact of the number of comprehension questions on metacomprehension accuracy has demonstrated a positive relationship between accuracy and the number of questions. Glenberg et al. (1987) found that the average point-biseral correlation significantly differed from zero when four comprehension questions were used to measure comprehension (r_{pb} = .21) but did not when a single comprehension question was used (r_{pb} = .07). In addition, using the materials developed by Glenberg and Epstein (1985), Weaver (1990) reported average Pearson correlations of r = .08, r = .19, and r = .32 for one, two and four-item measures respectively, with these differences between conditions found to be significant. Further, using simulations, Weaver (1990) showed that increasing the number of items to assess text comprehension could lead to a stronger alignment with metacomprehension judgements. This suggestion is corroborated by the results of a recent meta-analysis which has indicated that the number of comprehension questions has a small but significant effect on metacomprehension accuracy (Prinz et al., 2020a).

In addition to concerns about the number of comprehension questions, research suggests that the nature of comprehension questions can influence metacomprehension accuracy. The central proposition in prominent theories of text comprehension is that understanding involves the integration and synthesis of information to develop a network of interconnected meanings, with the role of inferences afforded special status within this framework as an often crucial component in successfully connecting information (Gernsbacher, 1990; Graesser et al., 1994; Kintsch, 1988; Trabasso & van den Broek, 1985; van den Broek et al., 1999; van Dijk & Kintsch, 1983). This mental network of the meaning of the text is often referred to as the text representation or situation-model (van Dijk & Kintsch, 1983; Zwaan, 2016; Zwaan & Radvansky, 1998). This framework of understanding reflects the fact that to comprehend a text means more than understanding any single word or sentence. Instead, to understand a text, you must draw meaning from, and link ideas across, multiple sentences, paragraphs, and background knowledge (Griffin, Wiley & Thiede, 2019; Kintsch, 1994).

Conceptualising text comprehension as the construction of an interconnected network of meanings arguably defines what we may consider to be appropriate targets for assessing comprehension of a text (Kintsch, 2012; O'Reilly et al., 2019). Targeting shallow forms of understanding when assessing text comprehension, therefore, may undermine the validity of estimated metacomprehension accuracy (Wiley et al., 2005). Research which has considered these concerns has emphasised the need to assess the situation-model level of understanding, using inference-based questions, to accurately measure text comprehension (Griffin, Wiley & Thiede, 2019). This approach identifies the appropriate assessment of comprehension as testing whether a connection between propositions has been made during reading. Examples of such assessments include assessing whether participants successfully connect information in different portions of the text; the presence of ideas formed on the basis of inferences drawn during reading; or the ability to make generative inferences from the situation-model (Jaeger & Wiley, 2014; Wiley et al., 2016). Despite the adoption of these methods to assess comprehension, however, low metacomprehension accuracy persists (Griffin, Wiley and Thiede, 2019; Jaeger & Wiley, 2014; Wiley et al., 2016, 2018).

Potentially, assumptions that participants may make about the nature of comprehension assessment could provide an explanation for failing to observe higher levels of metacomprehension accuracy, despite employing methods of assessing understanding which cohere with theoretical models of comprehension. In a series of experiments, Griffin, Wiley and Thiede (2019) demonstrated that participant's judgements of performance were consistently more predictive of scores on memory-for-detail than inference-based questions when no information about upcoming comprehension questions was provided. In contrast, when participants were informed that the comprehension questions would test their ability to connect pieces of information within the text, this pattern reversed. These findings are consistent with those of similar research (Thiede et al., 2011; Wiley et al., 2008) and are argued to reflect a default expectancy for comprehension questions which test surface-level details (Griffin, Wiley & Thiede, 2019). Therefore, even if performance measures are successful in assessing understanding at an appropriate depth, judgements of comprehension informed by test expectations may produce metacomprehension inaccuracy.

A further issue related to the nature of the questions is the central relevance of the information targeted in the question. It is acknowledged in theories of text comprehension that idea units, or propositions, within texts form a relational hierarchy of importance (Gernsbacher, 1990; Kintsch, 1988; Kintsch& van Dijk, 1978; Kuperberg & Jaeger, 2016; Yeari et al., 2017). Central propositions form the core structure of the representational model of the text, characterising the gist of the text (Kintsch & van Dijk, 1978). Not only are individuals able to identify and agree on idea units with greater central relevance to the semantic structure of the text (Brown & Smiley, 1977; Johnson, 1970), they also spend more time reading central propositions (Yeari et al., 2015) and recall them more frequently (Brown & Smiley, 1977; Yeari et al., 2017). Although the types and strength of coherence drawn from a text may vary considerably across reading episodes (van den Broek et al., 2011; van den Broek & Helder, 2017), successful understanding of a text is arguably characterised by the construction of a core network of meanings which accurately capture the central propositions. Consistent with the ‘good-enough’ model of comprehension (Ferrerira et al., 2002; Ferreira & Patson, 2007), therefore, some propositions and inferences may be unnecessary in successfully understanding the gist of a text.

Variability in the importance of propositions within a text may influence both metacognitive monitoring processes during reading and metacomprehension accuracy. For example, idea units which are more critical components within the semantic hierarchy of the text representation may be more closely monitored for coherence and may be more influential in evaluating whether understanding has been achieved. In support of this notion, main-idea violations in error-detection tasks have been found to be detected more easily than peripheral-detail inconsistencies (Baker, 1979; Yussen & Smith, 1990). This finding suggests that main-idea violations affect judgements of text coherence to a greater extent than incoherent propositions located at the periphery of the semantic structure of a text. If metacomprehension judgements likewise reflect a greater consideration of whether the most central and relevant propositions were understood, the alignment between perception and performance measures of comprehension could diverge where measures of comprehension do not target semantically central information. Consequently, considering the relative semantic importance of propositions within a text would, therefore, be important when designing appropriate measures of comprehension.

Unfortunately, it is often not clear whether the centrality of the idea or inference being assessed is robustly considered in metacomprehension accuracy research. While early research reported constructing inference-verification items which corresponded to the main ideas in stimulus texts, some inferences required greater elaboration than others making them harder to verify and, potentially, reducing metacomprehension accuracy (Glenberg & Epstein, 1985; Glenberg & Epstein, 1987; Glenberg et al., 1987). More recent research, which has emphasised the importance of testing understanding at the situation model level (Griffin, Wiley & Thiede, 2019; Wiley et al., 2016) has not explored the semantic centrality of assessed propositions and connective inferences. Comprehension test items may include inferences which correspond to peripheral propositions, with these connections featuring less strongly in the central propositions constructed from the text (Kintsch& van Dijk, 1978). If metacomprehension judgements are influenced more strongly by the established coherence of central propositions, performance measures which do not assess this will reduce observed metacomprehension accuracy (Wiley et al., 2005). However, as empirical research has not directly examined this issue, the potential role of semantic centrality remains uncertain.

1.4 Limitations of Research

It is clear from the research reviewed in section 1.3 that numerous variables may influence metacomprehension accuracy and, therefore, the utility of reader panels in evaluating health-related texts. However, the presence and direction of these effects is not consistent across studies. Some variables, such as background knowledge, appear only to have theoretical support, with empirical research indicating no benefit of greater knowledge on the ability to discriminate between texts which are more or less well understood (Jee et al., 2006; Griffin et al., 2009; Shanks & Serra, 2014). In addition, the observed relationships between variables and metacomprehension accuracy may not provide clear-cut interpretations. For example, whilst low ability readers may struggle with metacomprehension accuracy on texts which vary in difficulty, high reading ability does not ensure reliably high metacomprehension accuracy (Golke & Wittwer, 2017; Griffin et al., 2008; Lin et al., 2002; Maki et al., 2005; Ozuru et al., 2012). Also, the effect of some variables may be conditional on other factors. For example, the influence of the scope of the metacomprehension judgement, whether local or global, on metacomprehension accuracy appears inconsistent (Dunlosky, Rawson & McDonald, 2002; Nietfeld et al., 2005; Schraw, 1994), but may depend on a recall attempt prior to judgement (Dunlosky et al., 2005). Furthermore, while interventions appear to improve metacomprehension accuracy (Prinz et al., 2020b), this provides limited insight into the utility of judgements in the absence of such interventions, as is current practice in, for example, reader panels. The existing body of research, therefore, presents a challenge when considering the diagnostic validity of reader's judgements.

The considerable variation in methodologies employed within metacomprehension accuracy research could account for the observed inconsistencies in findings. As noted by Griffin et al. (2009), in any single study, there are a myriad of variables which may influence metacomprehension accuracy and obscure the influence of the variable under investigation. For example, Maki et al. (1990) and Maki and Serra (1992) elicited judgements at the level of paragraphs and found that predictions of performance produce higher metacomprehension accuracy than ease of comprehension judgements. In contrast, Ozuru et al. (2012) elicited judgements sentence-by-sentence and found that predictions of performance and ease of comprehension judgements produce similar levels of metacomprehension accuracy. As the scope of the judgement differed between these studies, possible variation in metacomprehension accuracy arising due to differences in the types of judgement may be obscured. Similarly, Jee et al. (2006) and Griffin et al. (2009) found that greater expertise benefits absolute metacomprehension accuracy of predictions of performance on short answer questions. Contrary to this, Shanks and Serra (2014) found that absolute accuracy of predictions of recalling single-sentence facts was harmed by greater expertise. While the effect of expertise appears inconsistent, this may occur due to differences in the appropriateness of the measure of text comprehension (Wiley et al., 2005).

Further complicating the above issues, individual studies are themselves not without limitations. For example, estimated metacomprehension accuracy may be undermined by the use of experimental measures which lack validity and reliability. For instance, to assess background knowledge, Jee et al. (2006) and Griffin et al. (2009) used participants' ratings of perceived expertise. Shanks & Serra (2014) also measured background knowledge using participants' perceptions of expertise, based on the order of texts ranked according to perceived topic knowledge. In both cases, not only are these measures vulnerable to inaccuracy, there is no guarantee that ratings or rankings reflect an equivalent degree of expertise between participants. Similarly, Maki et al.'s (2005) utilised verbal ability scores obtained from student records to capture reading skill which, if not reasonably contemporaneous with the research, could introduce bias in the measurement. Further, even in studies which use standardised reading ability tests, participants may be assigned to discrete categories for analytic convenience to capture varying levels of reading skill (Kurby et al., 2007, Ozuru et al., 2012), which itself can introduce bias in estimation (Altman & Royston, 2006; Fernandes et al., 2019).

More generally, the applicability of the findings of metacomprehension research in general, let alone to reader panels in health sentence, may be limited by the generalisability of the samples. Often, participant samples consist exclusively of university-level students and text samples include stimuli based on educational materials aimed at students. These may fail to generalise to both the wider population and to health-related texts. Metacognitive monitoring may be positively associated with academic achievement (see Tobias & Everson, 2009, for a discussion of this research), meaning that university students may have higher metacognitive skills than the general population (Garner, 1987). Potentially, therefore, metacomprehension accuracy may be lower across the population. Similarly, reading texts which are not novel in terms of their structure and complexity, as student participants often do when tested using stimuli based on educational material, contrasts with the likely lower exposure to these kinds of texts in the general adult population. Given potential differences in familiarity with instructional material and given that texts are themselves highly diverse (Wiley et al., 2005), it is not clear whether metacomprehension accuracy estimates derived from education texts generalize beyond an educational context.

Given the heterogeneity in findings and practice, alongside the homogeneity in samples of participants and texts, we are extremely limited in our ability to quantify the diagnostic validity of self-reported perceptions of comprehension of health-related information in the general population. It remains uncertain whether reader's judgements are likely to accurately evaluate whether key points of information will be comprehensible or are able to identify texts which engender incomplete understanding. It is evident, therefore, that additional research is required to provide healthcare practitioners with clear guidance on the utility of reader's judgement.

1.5 Research Aims

To further our understanding of metacomprehension judgements of health-related texts and to provide guidance on the utility of reader panellists' judgements, this research project aims to:

(1) Clarify the predictive association between metacomprehension judgements and comprehension of health-related texts in a sample of individuals not drawn from a student-only population.

(2) Examine whether individual differences in variables relevant to text comprehension and metacomprehension influence the predictive association between metacomprehension judgements and comprehension of health-related texts.

(3) Examine whether metacomprehension judgements can inform of overall text comprehension or of specific aspects of health-related texts.

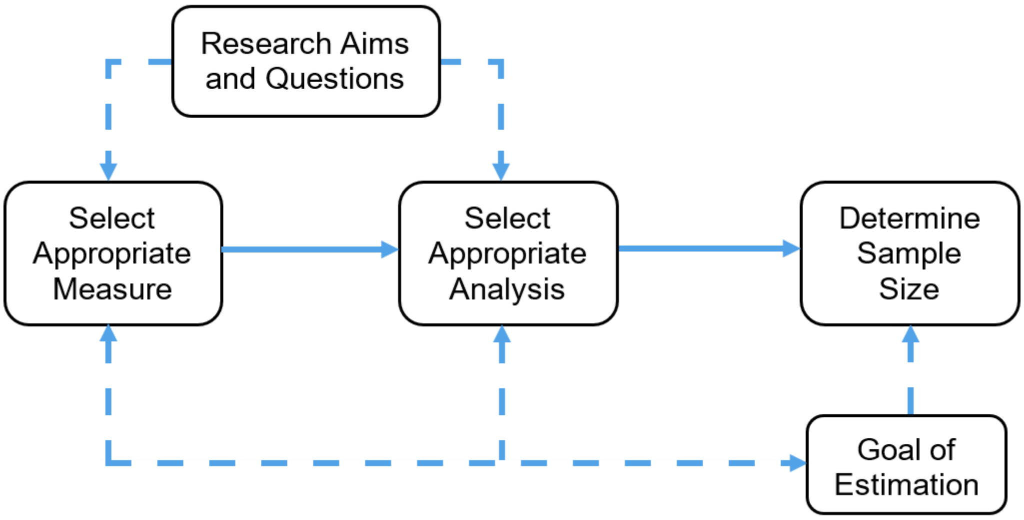
To address these research aims, several methodological factors must first be considered. Given their fundamental importance, these issues are explored in depth in Chapter 2, in the context of previous metacomprehension research.

2. Design and Inference Considerations

When designing quantitative empirical studies, multiple methodological concerns must be resolved to successfully address the research aims. Crucially, a researcher must select measures for the variables of interest which are valid and reliable. How phenomena are operationalised also, in part, determines the conclusions that may be drawn and, therefore, the selection of a measure must also be considered alongside the aims of the research (Schraw, 2009b). A researcher must also select an appropriate analytic approach to evaluate their observations. Whilst multiple analytic approaches may be statistically reasonable, the precise specification of an analysis will be influenced by the research questions. In addition, a researcher must determine the number of observations they must collect to be confident that the analytic approach they use will likely yield an estimation outcome permitting inferences about the phenomenon of interest (Johnson et al., 2015). How a researcher may define their target estimation outcome, or 'goal of estimation', influences the sample size required to achieve this (Kruschke, 2015). As the estimation goal operates relative to the chosen measure and analysis, these factors are also linked. The relationships between these three considerations and researcher-driven influences are illustrated in Figure 2.1.

Figure 2.1

Three Methodological Considerations in Study Design and Researcher-Driven Influences on Choices


Research Aims and Questions
Select Appropriate Measure
Select Appropriate Analysis
Determine Sample Size
Goal of Estimation

Note. Solid arrows illustrate the directionality of choices, dashed arrows indicate researcher-driven influences.

In the context of metacognitive judgements made by reader panellists, it is critical to consider the approach to measurement and analysis, in addition to the necessary sample size, since these factors significantly shape the insights the present research can provide. In the remainder of this chapter, previous research is discussed in relation to these three areas of concern, providing a clear foundation for the approaches adopted to address the research aims of this thesis.

2.1 Measuring Metacognitive Ability

The appropriate method to quantify an individual's metacognitive ability is a long-standing source of debate (Fleming & Lau, 2014; Higham & Higham, 2019). Over decades of research, multiple measures of metacognition have been proposed. According to Schraw et al. (2013), these measures can be classified according to the quantitative dimension of metacognitive ability which they capture. Five such categories of measures have been suggested: diagnostic efficiency, agreement, association, binary distance, and discrimination (Schraw et al., 2013). Within the field of metacomprehension, however, estimation approaches are typically categorised more coarsely as measuring either 'absolute' or 'relative' metacomprehension accuracy (Dunlosky & Lipko, 2007; Schraw, 2009a, 2009b; Thiede et al., 2019), though confidence bias has also been discussed as a distinct category (Griffin et al., 2009; Griffin, Mielicki & Wiley, 2019).

Measures of absolute accuracy capture the difference in magnitude between perception and performance scores (Bol & Hacker, 2012; Nelson, 1996). Typically, measures of absolute accuracy are calculated which provide signed or unsigned difference estimates of metacomprehension ability. The average signed difference between prediction and performance is referred to as an estimate of bias, expressing the overall tendency to over or underestimate performance. Alternatively, calculating the average unsigned difference between prediction and performance captures the magnitude of the deviations between prediction and performance. The calculation of absolute accuracy differs on several dimensions between studies, such as whether the metric is calculated using transformed prediction and performance scores, or itself undergoes transformation or standardisation. For example, in investigating the relationships between text difficulty, reading ability and metacomprehension accuracy (see section 1.3.2 for a discussion of findings), Golke and Wittwer (2017) and Maki et al. (2005) adopt different approaches to calculate bias. In Golke and Wittwer's (2017) study, bias was calculated as the difference between judgements and response accuracy, each scored from zero to six, yielding a bias score between -6 to +6. In contrast, Maki et al. (2005) calculated bias as a function of the total percentage discrepancy, using predictions of performance and summed performance scores each transformed into percentages.

Contrary to absolute accuracy, measures of relative accuracy capture the correlation between pairs of perception and performance scores (Bol & Hacker, 2012; Nelson, 1996). Typically, intra-individual correlations are calculated using judgements and summed performance across comprehension questions to provide a metric of relative accuracy which ranges from -1 to +1 (Griffin, Mielicki & Wiley, 2019). The use of relative measures is more prevalent in metacomprehension research than absolute measures, a preference rooted in Nelson's (1984) advocacy of Goodman-Kruskal's gamma correlation *G* (Goodman & Kruskal, 1954). As described in Chapter 1, *G* is a nonparametric measure of the probability of observing ordinal concordance between pairs of observations, capturing the tendency for higher perception scores to co-occur with higher performance scores (for a description of the calculation of *G*, see Gonzalez & Nelson, 1996). Consistent with Nelson's (1984) recommendations, metacomprehension studies predominantly calculate *G* (e.g., Dunlosky et al., 2005; Glenberg & Epstein, 1987; Lin et al., 2002; Maki et al., 2005; Margolin, 2013; Ozuru et al., 2012; Rawson et al., 2000; Thiede et al., 2010, 2011; Weaver, 1990; Wiley et al., 2018). Less frequently, parametric correlational measures are used to estimate relative accuracy, such as Pearson's *r* (e.g., Griffin et al., 2008, 2009; Griffin, Mielicki & Wiley, 2019; Linderholm et al., 2012; Thiede et al., 2003; Weaver, 1990; Wiley et al., 2018). Other correlational measures, such as the point-biserial correlation r_{pb} (e.g., Glenberg & Epstein, 1985; Glenberg et al., 1987) or Kendall's Tau-b (e.g., Maki & Berry, 1984), are far less common.

Despite the widespread use of absolute and relative measures of accuracy in metacomprehension studies, several problems with these measures have previously been identified. Early criticism of measures of absolute accuracy highlighted the vulnerability of these measures to non-metacognitive influences (Nelson, 1984). Specifically, performance levels can vary, due to factors such as test difficulty, while underlying metacomprehension ability remains constant (Griffin, Mielicki & Wiley, 2019). Estimated differences in measures of absolute accuracy may not, therefore, reflect differences in metacomprehension ability (Nelson, 1984). In addition, researchers have argued that discrepancy and change scores can have low reliability (Cronbach & Furby, 1970; Hattie, 2013; Lord, 1956), with reliability dependent on the variability in judgement and performance scores and their underlying association (Trafimow, 2015). More generally, the inability of measures of absolute accuracy to go beyond ordinal statements of difference and the consequent limited range of research questions which may be meaningfully addressed has been discussed (Benjamin & Diaz, 2008). Griffin, Mielicki and Wiley (2019) have likewise argued that measures of absolute accuracy provide limited theoretical insight within the field of metacomprehension.

With respect to relative accuracy, researchers have similarly highlighted problems with the most frequently used measure *G*. Benjamin and Diaz (2008) found a nonlinear relationship between estimated *G* and simulated underlying metacomprehension ability when *G* is close to the boundaries of -1 and +1, indicating a limited capacity for *G* to capture extreme values of metacomprehension ability. In addition to this, Masson & Rotello (2009) demonstrated that the exclusion of tied judgement-performance paired scores, as is advised in the calculation of *G* (Gonzalez & Nelson, 1996), leads to systematic bias in estimation, with the extent of the overestimation of *G* dependant on the strength of judgement bias. Overestimation of *G* also occurs if fewer response options for perception judgements are provided, due to an increase in the number of ties (Higham & Higham, 2019). It has further been suggested that *G* may only be an unbiased estimator of true underlying discrimination in asymptotically large samples, while smaller samples are likely to lead to distributions of values which are abnormal, skewed and highly variable (Goodman & Kruskall, 1963, 1972; Masson & Rotello, 2009). Consistent with this assertion, these characteristics are observed in empirical distributions of *G* in metacomprehension studies (Griffin, Mielicki & Wiley, 2019).

Researchers have also identified issues with the use of other measures of relative accuracy, including Pearson's *r*. Consistent with the findings regarding *G*, Benjamin and Diaz (2008) also found a nonlinear relationship between estimated *r* and simulated underlying metacognitive ability, indicating a general limitation of measures with hard boundaries to accurately express the extremes of metacognitive ability. In addition, the assumptions of parametric measures of association may not be reasonably satisfied, such as linearity and interval-level measurement scales, making them unsuitable to estimate metacognitive ability (Nelson, 1984). Further, Nelson and Narens (1990) have contended that effectively handling tied observations is problematic using measures such as Pearson's *r* and Kendall's Tau, as these approaches do not prevent tied observations of judgement and performance scores, which may be considered theoretically irrelevant, from contributing to estimated metacognitive ability (Nelson, 1984). Researchers have also raised concerns, more generally, regarding the properties of parametric measures, in terms of bias and distribution in the context of limited numbers of observations, given the lack of research on this scenario (Schraw, 2009a).

Given the problems identified with measures of absolute and relative accuracy, researchers have advocated for the use of measures which are based on signal detection theory (SDT). The SDT framework posits that behaviour in a binary discrimination task is the result of the classification of a continuous, latent distribution of an individual's subjective experience of signal intensity (or 'evidence distribution'), consisting of the combination of a fixed signal strength and random noise, according to a decision threshold (Green & Swets, 1966; Wixted, 2020). Within the SDT framework, a distinction is made between two types of discrimination task, termed 'type 1' and 'type 2' tasks (Clark et al., 1959). The former involves discriminating between objective states of the world which are independent of the observer, while the latter involves discriminating between the states of one’s own decisions (Galvin et al., 2003). For example, in the context of metamemory research, participants may be asked to provide a feeling-of-knowing judgement, capturing their certainty in knowing the answer to a stimulus on an upcoming trial, followed by a binary verification response on the trial itself, and a judgement of their confidence in the accuracy of their verification response (e.g., Mazancieux et al., 2020). The verification response to the stimulus is a type 1 task, while the judgements of knowing and confidence are type 2 tasks.

The SDT framework provides several measures of an individual's sensitivity to stimulus intensities, capturing the ability to discriminate between signal states with accuracy (Verde et al., 2006). Researchers have argued that these measures are preferable to quantify an individual's ability to discriminate between stimulus states, as they are free from the bias inherent in measures such as *G* (Benjamin & Diaz, 2008; Fleming & Lau, 2014; Masson & Rotello, 2009). However, researchers have suggested that type 2 SDT measures may be confounded by metacognitive bias, as the assumption that judgements result from the classification of a normally distributed underlying evidence distribution with equal variance may be violated in type 2 tasks (Fleming & Lau, 2014). Instead, the SDT-based measure meta-*d'* has been argued to appropriately quantify metacognitive ability free of response bias (Fleming, 2017; Maniscalco & Lau, 2012, 2014). More recently, receiver operator characteristic (ROC) curves, an analytical approach within the SDT framework (Wixted, 2020), have been used to provide an approach to estimating *G* which produces estimates closer to simulated underlying metacomprehension ability, with lower variability, than the traditional approach to estimating *G* (Higham & Higham, 2019).

Despite SDT measures being commonplace in metacognitive discrimination tasks, a small proportion of metacomprehension studies report SDT measures and, typically, not as the main estimate of interest (e.g., Wiley et al., 2018). The limited adoption of SDT measures in metacomprehension studies is, in part, due to critiques which have argued that the statistical assumptions of SDT measures cannot be verified and may not hold across metacognitive tasks (Nelson, 1984, 1986; Nelson & Narens 1990). For example, it has been suggested that assuming that i) the underlying evidence distributions are normally distributed with equal variance and that ii) the threshold for discriminating between evidence states is invariant across trials, may not be empirically or theoretically substantiated (Nelson, 1984; 1986). When neither the shape of the evidence distribution nor the stability of the discrimination threshold is known, differences in estimated SDT measures may correspond to non-metacognitive influences, such as response bias and performance accuracy (Nelson, 1984, 1986; Nelson & Narens 1990).

More pragmatically, the use of SDT measures in metacomprehension research is constrained by experimental design. For example, metacognitive judgements correspond to multiple items in metacomprehension research (Griffin, Mielicki & Wiley, 2019), whereas in type 2 SDT tasks judgements are collected at the level of individual test items (Galvin et al., 2003). In addition, to calculate SDT measures, performance must be assessed using a two-alternative forced-choice task (Fleming, 2017). In the context of comprehension, this restricts assessment to sentence and inference verification tasks (e.g., Glenberg and Epstein, 1985), yet multiple-choice questions are preferred in metacomprehension research (Prinz et al., 2020a). This limitation has theoretical implications, since the response format may influence which aspects of comprehension processes are assessed and, therefore, how metacomprehension accuracy itself is operationalised (Ozuru et al., 2013; Prinz et al., 2020a). Adjacent to these concerns, researchers have also questioned the feasibility of collecting sufficient observations to accurately estimate SDT measures, or non-parametric SDT alternatives, within metacognitive tasks (Griffin, Mielicki & Wiley, 2019; Nelson, 1984).

An alternative approach, ameliorating the limited capacity to handle complex designs within the SDT framework, is the use of regression models to estimate a model parameter representing metacognitive ability. For example, metacognitive studies of perception and memory have used the estimated slope coefficients of regression models to directly provide individual-level estimate of metacognitive ability (e.g., Faivre et al., 2020; Mazancieux et al., 2020; Sandberg et al., 2010, 2013). However, Rausch and Zehetleitner (2017) have demonstrated that regression slope coefficients are sensitive to statistical assumptions concerning the distribution of evidence underlying stimulus discrimination decisions, whether and to what extent information is shared in forming judgements in type 1 and type 2 tasks, and response bias. More recently, Kristensen et al. (2020) demonstrated that the proportional odds model, a form of ordinal regression, extended to incorporate a latent variable, could be used to express the SDT model with meta-*d'* explicitly modelled as a parameter within the model. Established measures drawn from the SDT framework may, therefore, be preferable to slope coefficients from regression models, conditional on the suitability of the task design to the SDT framework.

Complicating matters, however, regardless of which metric is used to measure metacognitive ability, the resulting estimate may be influenced by how we study metacognition. Considering the impact of study design, Vuorre and Metcalfe (2022) demonstrate that, in tasks which permit guessing, measures of metacognitive ability, including correlational and SDT measures, are unavoidably linked to task performance. Estimated metacognitive ability was shown to become increasingly negatively biased as performance ability decreased, as task responses increasingly relied upon guessing (Vuorre & Metcalfe, 2022). This effect is greatest in tasks with very few alternative response options, such as two-alterative forced choice tasks, due to the greater potential contribution of guessing to observed performance scores (Schwartz & Metcalfe, 1994; Vuorre & Metcalfe, 2022). Vuorre & Metcalfe's (2022) findings suggests that open-ended questions may reduce the impact of response bias on estimated metacomprehension accuracy. However, as noted, since the choice of response format may affect how comprehension processes are assessed (Ozuru et al., 2013), metacomprehension accuracy may also be influenced by the response format (Prinz et al., 2020a).

More broadly, estimates of metacomprehension accuracy may be undermined by insufficiently identifying and controlling confounding variables. Paulewicz et al. (2020) highlights the paucity of understanding of metacognition in connecting stimulus processing to responses and metacognitive decisions. This lack of knowledge requires that researchers consider multiple causal relationships to describe behaviour in metacognitive tasks (Paulewicz et al., 2020). In the absence of strong causal assumptions or study designs which aim to break potential confounding paths, Paulewicz et al. (2020) argues that any measure of metacognitive ability which seeks to remove bias arising from the dependence between the cognitive mechanisms underlying observed perception and performance, without explicitly conditioning on such influences, fails to sufficiently address this issue (Pearl, 2000). As a result, measures of metacomprehension accuracy, including both correlational and SDT measures, may be biased by unidentified confounding variables (Paulewicz et al., 2020). This issue is of particular concern where observational measures of metacomprehension accuracy are used to substantiate a causal account of metacomprehension judgements.

2.1.1 Selecting a Measure of Metacomprehension Accuracy

Given the discussion in section 2.1, it is evident that the range of existing measures of metacomprehension accuracy are not without limitations. Absolute measures can suffer from low reliability and may provide little insight into metacomprehension processes (Griffin, Mielicki and Wiley, 2019; Hattie, 2013). Correlational measures may provide biased estimates and fail to fully capture variability metacomprehension ability (Benjamin & Diaz, 2008; Higham & Higham, 2019). Further, for parametric correlational measures, the underlying assumptions are likely not fully met in metacomprehension research designs (Nelson, 1984). Measures from the SDT framework lack applicability to standard metacomprehension study designs (Griffin, Mielicki & Wiley, 2019), and altering the task demands to be consistent with the SDT framework increases the risk of introducing bias due to correct guessing (Vuorre & Metcalfe, 2022). Similarly, regression slope coefficients may be sensitive to non-metacognitive influences, such as response bias (Rausch & Zehetleitner, 2017). Further, interpreting any of these measures as an accurate metric of an individual's latent metacognitive ability is problematic due to unaddressed confounding variables (Paulewicz et al., 2020). It is perhaps unsurprising, then, that researchers have conceded that no perfect measure exists and that the idea of isolating elements of metacognition might reasonably be abandoned (Higham & Higham, 2019; Paulewicz et al., 2020).

In the context of judgements of the comprehensibility of health-related information, the key question then is: how should we seek to meaningfully quantify the utility of subjective evaluation of text comprehensibility? Judgements of comprehension are used to evaluate whether health information is likely to engender understanding in patients, based on the assumption that reader panellists' judgements are predictive of their actual understanding of the text. The assumption of a predictive relationship between perceived and actual comprehension indicates that it would be appropriate to estimate the systematic association between these variables across texts. Arguably, using regression coefficients to quantify the relationship between measures of perceived and assessed comprehension is reasonable in the present research, as this approach to measurement would provide insight into whether or not metacomprehension judgements show predictive validity.

Importantly, in selecting a regression-based approach to measure the predictive relationship between perceived and assessed comprehension on health-related texts, a fundamental concern must be addressed. Specifically, we should consider the extent to which such an analysis is informative of an individual's underlying metacomprehension ability (Paulewicz et al., 2020). Treating regression slopes as an accurate measure of underlying metacomprehension ability is problematic, given the potential influence of non-metacognitive factors (Rausch & Zehetleitner, 2017; Vuorre & Metcalfe, 2022). In this thesis, therefore, individual-level coefficients will not be used as the basis for drawing strong inferences concerning latent metacomprehension ability. Nonetheless, since regression slopes (or curves) are sensitive to differences in underlying metacomprehension ability (Rausch & Zehetleitner, 2017), it is reasonable to explore how variables considered theoretically relevant to metacomprehension ability may impact on the predictive association between perception and performance measures. Further, variability in the predictive relationship permits speculation concerning underlying metacognitive ability, though firm conclusions may not be justified (Rausch & Zehetleitner, 2017).

2.2 Analysing Differences in Metacomprehension Accuracy

To evaluate whether an experimental manipulation affects metacomprehension accuracy, the dominant approach within metacomprehension research involves a two-step analysis procedure. The first step is to estimate the chosen metric of metacomprehension accuracy using observations of judgements and performance from each individual. Secondly, these individual-level estimates are used as the outcome variable in a statistical analysis of between-group differences, such as a t-test, analysis of variance or regression. For example, Lin et al. (2002) first calculated *G* for each individual, followed by an analysis of variance to statistically evaluate whether the average *G* differed between age groups and levels of text difficulty. Similarly, Kwon and Linderholm (2014) first calculated a metric of absolute accuracy for each participant, before using these values as the outcome of a regression analysis to examine whether the discrepancy between an individual's perception and the reality of their reading ability predicts absolute accuracy.

Despite the two-step analysis procedure remaining the main approach for almost four decades within metacomprehension accuracy research (e.g., Dunlosky et al., 2005; Glenberg & Epstein, 1985; Glenberg et al., 1987; Griffin et al., 2008, Griffin, Wiley & Thiede, 2019; Kwon & Linderholm, 2014; Lefevre & Lories, 2004; Lin et al., 2002; Linderholm et al., 2012; Ozuru et al., 2012; Rawson et al., 2000; Thiede & Anderson, 2003; Thiede et al., 2003, 2010, 2011; Vossing, 2019; Weaver, 1990; Wiley et al., 2018), critical limitations of this approach can be identified. Firstly, using individual-level metrics of metacomprehension accuracy as an outcome variable discards all information about the uncertainty associated with the estimates. For example, Thiede et al. (2003) used six pairs of judgements and summed performance scores to estimate *G* for each participant, before using these estimates as the dependent variable in an analysis of variance to examine the influence of keyword generation on metacomprehension accuracy. Due to the small number of observations per participants, individual-level measures of metacomprehension are likely to be associated with high levels of uncertainty, if this may even be quantified for non-parametric measures given the volume of data (van der Ark & van Aert, 2014). Erroneously treating individual-level estimates of metacomprehension accuracy as error-free in the second step may lead to unreliable results and spurious conclusions (Boehm et al., 2018; Kristensen et al., 2020).

In addition, the two-step analysis approach does not partition variance associated with non-independent elements within metacomprehension studies. As observations of judgements and performance are obtained using the same texts and questions, participants responses can be considered nested within these stimuli. For example, in Rawson et al.'s (2000) first experiment investigating the influence of rereading on metacomprehension accuracy, participants provided judgements on the same six texts and answered the same six comprehension questions per text. As a result of this design, responses obtained on each text will share variance across participants. In the first step, in calculating a measure of metacomprehension accuracy, performance on each text is calculated as the sum total of correct responses to comprehension questions per text. In doing so, responses to each question are considered equitable, despite likely variability at the question-level across texts (Rouder & Haaf, 2019). In the second step, estimates of metacomprehension accuracy are treated as independent. However, variance is shared across participants as a result of the use of the same texts and questions. Failing to explicitly account for such hierarchical structured variance can lead to bias in estimation and misleading findings (Boehm et al., 2018; Baayen et al., 2008).

Furthermore, individual-level estimates of metacomprehension accuracy may not fully satisfy the requirements of the statistical test applied in the second step. In addition to violating the assumed independence of observations, assuming that the residual variance in participant-level measures is normally distributed is questionable, given the distributions that are empirically observed (Griffin, Mielicki & Wiley, 2019). Moreover, even if approximate normality is observed at the sample level, a lack of knowledge of the distributional properties of metacomprehension accuracy measures at the population level persists (Schraw, 2009a). For measures of metacomprehension accuracy which yield bounded values, the use of tests which assume a Gaussian distribution may not be theoretically justified, rendering the analyses less meaningful. Furthermore, with respect to the independent variables of interest, the selection of statistical tests which require variables to be discretised, such as the use of analysis of variance, can produce biased estimates of between-group differences (Altman & Royston, 2006; Fernandes et al., 2019).

2.2.1 Selecting an Analytic Approach

It is evident, given the issues surrounding the standard two-step approach identified in section 2.2, that an alternative approach is required to robustly evaluate differences in estimated metacomprehension accuracy. Fortunately, complementary to the selection of regression coefficients to express the predictive association between perceived and assessed comprehension (as described in section 2.1.1), regression models offer a suitable alternative approach. Specifically, logistic multilevel regression provides a flexible and cohesive framework which is, therefore, well-justified in the present research. Logistic regression is used to analyse binary response data, with the contribution of predictor variables on the probability associated with the binary outcomes modelled using a logistic link function (Agresti, 2002; Hilbe, 2009). Multilevel regression introduces additional parameters into the regression model to capture the hierarchical structure of variance generated through the experimental design (Baayen et al., 2008; Gelman & Hill, 2007). In multilevel logistic regression, therefore, the probability associated with the binary outcomes and the effects of predictor variables can be modelled as varying within groups, such as participants and stimuli, often referred to as 'random intercepts' and 'random slopes', respectively (Snijders & Bosker, 2011).

The limitations of the two-stage analysis can be fully addressed by using multilevel logistic regression to analyse differences in metacomprehension accuracy. Participants' text-level metacomprehension judgements can be entered as a predictor of the accuracy of question-level responses to calculate the average estimated association between judgements and performance, while individual deviations from this average are estimated at the participant-level. Concomitantly, non-independent variability generated by repeated observations of texts and comprehension questions can be incorporated. This resolves the violations of assumed independence of observations, increasing the accuracy of estimated effects and standard errors, whilst simultaneously improving efficiency in the estimation of individual-level variances via partial pooling (Gelman & Hill, 2007). In addition, the logistic model does not require that the residuals for the response variable are normally distributed, and observations of judgements and performance are not assumed to be interval-level data. Further, multiple variables can be included as predictors, with no discretisation required to test for differences (Kristensen et al., 2020). Overall, therefore, this approach provides a more efficient way to handle the uncertainty in observations, with assumptions which are more easily theoretically justified.

To further improve on the capacity to quantify and express uncertainty in estimated effects, regression analyses can be conducted within a Bayesian estimation framework. In the Bayesian approach to estimation, parameters are estimated using a combination of prior knowledge and information in the data, resulting in a posterior distribution which expresses the plausible values for the parameter (Kruschke, 2015; van de Schoot et al., 2021). According to this framework, therefore, a model parameter is represented as a distribution of probable values, rather than as a fixed-point value (Kruschke & Liddell, 2018; van de Schoot et al., 2021). Conceptualising parameters as distributional has been argued to provide a more informative and intuitive approach to estimation, as the probability associated with values can be directly quantified using the posterior distribution (Kruschke, 2013; Kruschke & Liddell, 2018). In contrast, frequentist confidence intervals provide no information about the relative probability of values contained within the interval for an effect. Given the additional ability to quantify the uncertainty in the magnitude of estimated effects, metacomprehension accuracy will be analysed using Bayesian multilevel logistic regression in this thesis.

2.3 Determining the Required Sample Size

Ensuing that a study will generate data which are capable of addressing the research questions is a crucial consideration (Johnson et al., 2015). Based on Neyman and Pearson's (1928a, 1928b, 1933) tests of statistical hypotheses, researchers currently tend to operate within a null hypothesis significance testing (NHST) framework (Calin-Jageman & Cumming, 2019). According to this framework, investigating whether an experimental manipulation results in a difference in a variable of interest (i.e., an effect) involves statistically evaluating whether or not the observations are consistent with there being no difference between experimental conditions. For example, for a two-tailed test of significance, the null hypothesis states that the effect of interest is zero and the alternative hypothesis states that the effect is not equal to zero. Evidence against the null hypothesis is generally considered to be obtained when an effect estimate is observed with an associated *p*-value of less than .05 (Kruschke & Liddell, 2018). Statistical significance, given this *p*-value, corresponds to an expectation that estimates for an effect which are equal to, or more extreme, than the observed effect would be obtained on 5% of the hypothetical occasions the study is conducted, given that the null hypothesis is true. According to this framework, observing a *p*-value of < .05 is considered evidence which permits the researcher to make inferences about the existence of a non-zero difference at the population-level (Kruschke & Liddell, 2018; Calin-Jageman & Cumming, 2019).

Since observing a statistically significant effect, and thereby rejecting the null hypothesis, is the goal of estimation in the majority of studies, researchers typically evaluate the sample size required to address their research questions by using a statistical power analysis. A statistical power analysis evaluates the *a priori* probability of correctly rejecting the null hypothesis, when the alternative hypothesis is true, for a prospective study (Cohen, 1962, 1988, 1992a, 1992b). A study with adequate statistical power to achieve this goal of estimation is typically defined as providing evidence against the null hypothesis, demonstrating an effect which is significantly different from zero, assuming it exists, on 80% of the (hypothetical) occasions that the study is conducted (Cohen, 1965, 1992a). Minimally, to conduct a statistical power analysis, a researcher must provide values for the true difference and the variance at the population level (Perugini et al., 2018). However, additional parameter values are required if a researcher wishes to conduct more complex analyses of their data, such as the intra-class correlation coefficient for hierarchical analyses (Hedges & Rhoads, 2010). Differences in the calculation of power occur because a statistical power analysis computes the probability of rejecting the null hypothesis for a given type of statistical test (i.e., the power of the test).

For metacomprehension studies with a standard design and using the typical approach to measurement and analysis, a statistical power analysis would identify the required sample size likely to yield a statistically significant difference between experimental conditions in the average value of whichever metric of metacognitive ability is calculated. Despite the capacity to formally evaluate the necessary sample size, the vast majority of metacomprehension studies do not discuss power concerns, though usage of power analyses to a priori determine the required sample size has increased in recent years (e.g., Griffin, Wiley & Thiede, 2019;

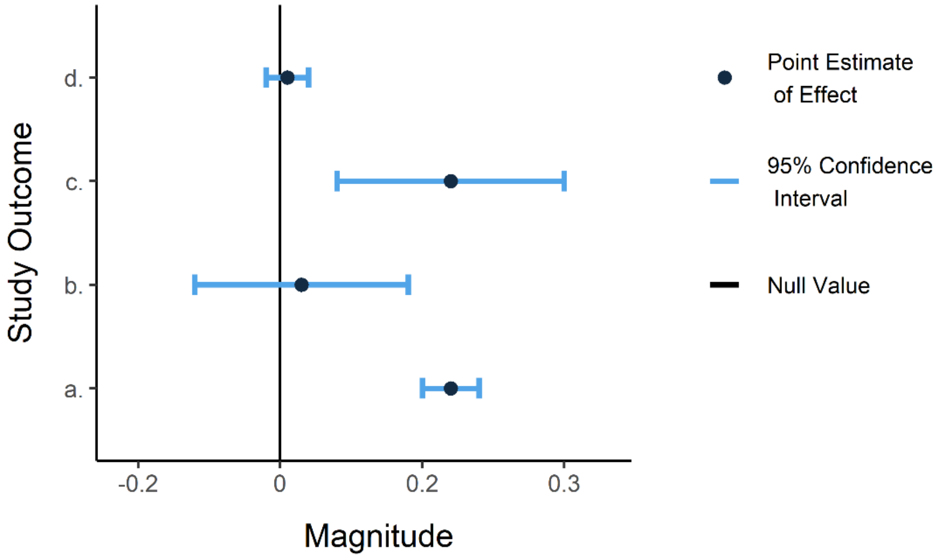
Madison & Fulton, 2022; Thiede et al., 2022; Vuorre & Metcalfe, 2022; Yang et al., 2018). However, even when statistical power analyses are conducted in metacomprehension research, the required sample size may be underestimated. In analysing the power of the typical two-step analytic approach, the quantity of data required to suitably estimate the individual-level measure of metacomprehension accuracy is not formally considered. For example, if metacomprehension accuracy is quantified using Pearson's *r*, with six pairs of perception and performance scores per participant, the true coefficient must be at least approximately .91 to yield an 80% probability of concluding the coefficient is reliably not in fact zero. Given that only the variability between and not within individuals is considered in the calculation of statistical power analysis for metacomprehension studies, it is likely that variance is underestimated, resulting in lower statistical power for a given sample size.

Researchers have repeatedly attempted to draw attention to the complications of conducting studies with inadequate statistical power when using the NHST framework for inference (Anderson et al., 2017; Button et al., 2013; Cohen, 1962, 1990; Maxwell, 2004; Smaldino & McElreath, 2016; Vankov et al., 2014; Wicherts et al., 2016). Conducting studies with low power, in the context of NHST, leaves the researcher vulnerable to obtaining data which are insufficient to reject the null hypothesis, despite the existence of a true effect at the population-level. Further, while statistical significance may be achieved, when power is low, the effect may be poorly estimated. In fact, conditional on statistical significance, in an underpowered study, effect estimates which are statistically significant are more likely to be inaccurate than accurate (Gelman & Carlin, 2014). Significant effects may include those in the wrong direction to the true effect (a sign error) or effects which are grossly exaggerated (a magnitude error). Moreover, even when the typically desired 80% power is achieved, for statistically significant effect estimates, the expectation of the observed effect is an overestimate of the true effect (Gelman & Carlin, 2014). Coupled with the bias towards publishing significant findings which elevates the false discovery rate (Sterling et al., 1995), serious questions can be raised about the accuracy of published research.

Even where the required sample size is identified and a suitably powered study is conducted, researchers have argued that the goal of estimation in the NHST framework can limits our capacity to learn about the effect of interest (Cumming, 2014; Kruschke, 2013). Figure 2.2 demonstrates four hypothetical study outcomes given an NHST approach to inference. In outcomes a. and c. in Figure 2.2, statistical significance is achieved but the standard error of the estimate may comparatively be low (a.) or high (c.). The NHST framework treats these study outcomes as equivalent as both result in successful rejection of the null hypothesis. However, in terms of the estimation of the effect, in outcome c. we learn comparatively less about the magnitude of the effect.

Figure 2.2

Effect Point Estimates and 95% Confidence Intervals Obtained from Hypothetical Studies



Similarly, while statistical significance is not observed in outcomes b. and d. in Figure 2.2, a lower standard error of the estimate is observed for the latter outcome. According to an NHST framework, what can be inferred about the effect at the population-level is the same given these outcomes. Yet, considering the estimation of the effect in outcome d., while a researcher cannot make claims about the directionality or existence of the effect, the study nonetheless provides confidence in a narrower range of values which the effect may take. Essentially, the binary decision to reject or fail to reject the null hypothesis is a crude method to evaluate what may be learned from a study (Calin-Jageman & Cumming, 2019; Kruscke, 2013) and may limit theoretical progress (Meehl, 1990).

The NHST framework has remained the dominant approach to inference for decades (Chavalarias et al., 2016), mandating that a researcher's goal of estimation is the correct rejection of the null hypothesis. Despite this, researchers typically want to achieve more than the blunt categorisation of whether or not an effect exists (Cumming, 2012, 2014; Kruschke, 2013; Trafimow, 2019). Principally, researchers are often motivated to learn about the likely magnitude of an effect and to quantify their confidence in the range of values which an effect may take. In advocating for a shift from the NHST framework, Cumming (2014) argues that adopting an estimation-centred approach to inference, focusing on the confidence interval for an effect, would improve both the integrity and progress of research. Frequentist 95% confidence intervals represent the range of values that, if we repeated the experiment infinitely, contain the true parameter value on 95% of occasions (Neyman, 1937). While these intervals do not express the probability of the true effect value being captured within a single set of interval bounds, in contrast to *p*-values, confidence intervals provide considerably more information about an effect and readily allow the researcher to interpret the uncertainty associated with the estimated effect (Cumming, 2014; Hoekstra et al., 2014).

Further, adopting an interval-based approach to inference within a Bayesian estimation framework has been argued to provide additional benefits (Kruschke, 2013; Kruschke & Liddell, 2018). Bayesian credible intervals have been argued to be well-suited to quantify a researcher's confidence in the range of values an effect may take (Kruschke, 2015). A Bayesian credible interval for an effect is calculated using the posterior distribution, capturing the range of values with a selected total probability density, such as 95%. Credible intervals are commonly calculated as central intervals or highest density intervals (Kruschke, 2015; Lambert, 2018). Central intervals, or equal-tailed intervals, are calculated using quantiles to exclude an equal probability density in the tails of the posterior distribution, whereas highest density intervals capture the range of values with the highest probability density for which the total density integrates to some desired probability value. For unimodal, symmetric, posterior distributions, these intervals are equivalent. In contrast to frequentist confidence intervals, credible intervals are able to quantify the relative probability of values contained within the interval, providing greater insight and a more natural interpretation of uncertainty (Kruschke & Liddell, 2018; Morey et al., 2016).

As illustrated in Figure 2.2, obtaining confidence intervals or credible intervals which allow the researcher to substantively learn about an effect can provide greater insight than the binary decision of whether or not to reject the null hypothesis. The width of an effect interval conveys what may be learned about the effect, with narrower intervals reflecting a more concise set of values with which we can express our confidence of containing the true effect (Cumming, 2014; Maxwell et al., 2008). It is reasonable to expect, therefore, that a researcher may be motivated to obtain a narrow interval for an effect to address their research question (Calin-Jageman & Cumming, 2019). This concern constitutes an alternative goal of estimation: achieving precision in estimation. To render this goal likely to be achieved, the sample size required must be identified. As a statistical power analysis is only equipped to evaluate the probability of rejecting the null hypothesis, researchers must use a different method to quantify the probability of achieving the goal of precision in estimation.

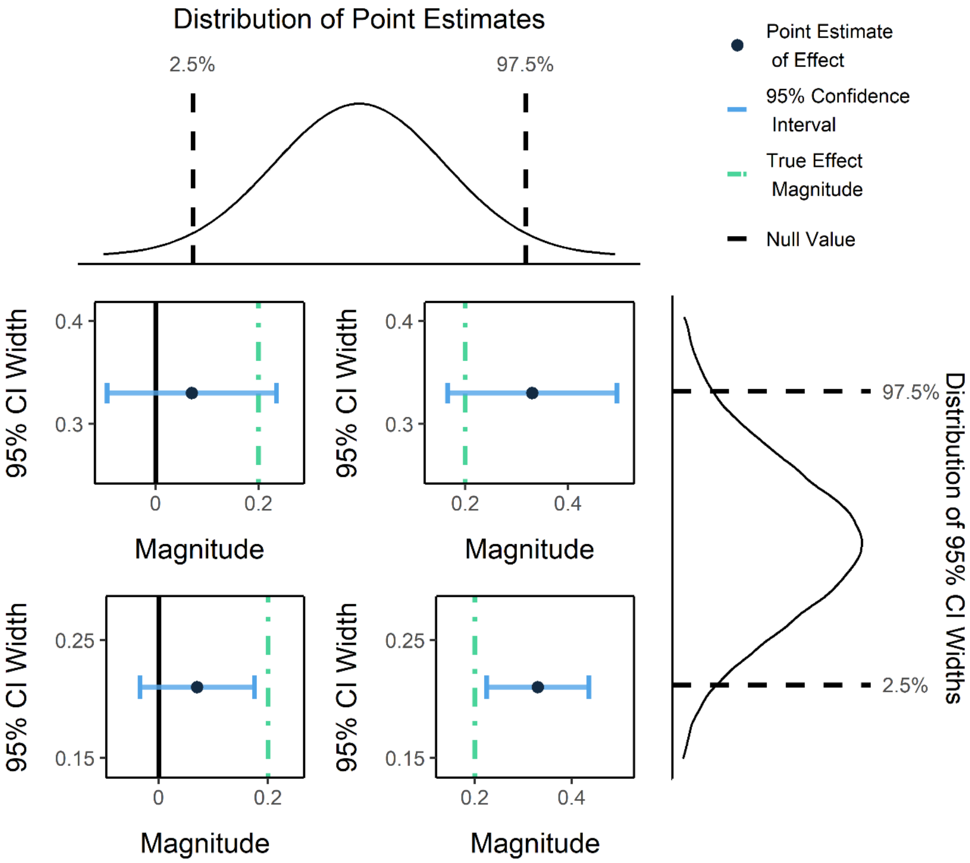
Precision analysis, also referred to as 'accuracy in parameter estimation' (AIPE; Maxwell et al., 2008), examines the estimated uncertainty associated with effect estimates for a given study design and analysis (Kelley et al., 2003; Johnson et al., 2015; Landau & Stahl, 2013; Rothman & Greenland, 2018). To evaluate the a priori probability of achieving precision in estimation, the researcher must select a target width *w* for the effect interval, which expresses the level of precision they consider to be acceptable (Kelley et al., 2003; Maxwell et al., 2008). As the probability of achieving precision in estimation is determined by the population-level variability and sample size, *w* can be selected without consideration of the magnitude of the population-level effect (Kelley et al., 2003, Maxwell et al., 2008). For simple study designs and analyses, therefore, only the selection of *w* and the population-level variance is required to calculate the required sample size which provides a sufficient probability of achieving precision in estimation (Kelley et al., 2003).

Whilst precision analysis is indifferent to both effect values and the null value (Kelley et al., 2003), however, it is important to consider these values when identifying an acceptable level of precision for a prospective study if the magnitude of the effect is of interest to the researcher (Maxwell et al., 2008). Yet the lack of a formal method to robustly evaluate this leaves the researcher vulnerable to failing to fully consider the implications of a chosen *w*. As a result, undesirable study outcomes may occur despite achieving the target level of precision in estimation. To illustrate, we may consider the potential outcomes of a study estimating the difference between two groups in the mean value of a variable of interest when the target precision is 'low'. For this example, assume that in both groups, observations are independent and the variance is normally distributed, with standard deviation $\sigma = 0.2$. Further, the true difference in means between the groups is 0.2. To aid understanding, we could consider the two groups correspond to a control group and a situation-model based intervention group, with the intervention leading to a 0.2 increase in the measure of metacomprehension accuracy (Prinz et al., 2020b). A target width *w* for the 95% confidence intervals is selected as 0.3. Based on *w* and σ, a required sample size of 36 participants provides an 80% probability of observing confidence intervals with widths of ≤ 0.3.

Given the effect magnitude, population variance, and sample size, the sampling distributions of both the effect estimates and the 95% confidence interval widths can be defined. To consider the potential impacts of *w* = 0.3, these distributions and four possible study outcomes are shown in Figure 2.3. In this scenario, we can expect to observe an effect magnitude equal to, or more extreme than, the point estimates plotted (≤ 0.07 and ≥ 0.33) on 5% of the hypothetical occasions the study is conducted. The study outcome panels on the left and right columns show the lower and upper 2.5% of point estimate magnitudes, respectively. Similarly, we can expect to observe a 95% confidence interval with a width equal to, or more extreme than, the plotted confidence intervals (≤ 0.22 and ≥ 0.32) on 5% of the hypothetical occasions the study is conducted. The panels on the top and bottom rows show the lower and upper 2.5% of confidence interval widths, respectively.

Figure 2.3

Distributions of Point Estimates and 95% Confidence Interval Widths for the Mean Difference in a Hypothetical Study, with Four Possible Estimation Outcomes



Note. The dashed lines at 2.5% and 97.5% on the distribution of point estimates correspond to the location of the point estimates, shown as dots on the horizontal axis in each panel. The dashed lines on the distribution of 95% confidence interval widths correspond to the magnitude of widths labelled on the vertical axis, with the range of effect magnitudes contained within these intervals shown on the horizontal axis. CI = confidence interval.

The bottom row of Figure 2.3 demonstrates how, despite achieving the target level of precision in estimation (observed width is ≤ 0.3), the 95% confidence interval can contain the null value and fail to include the true effect (on 5% of occasions). Further, the top row of Figure 2.3 indicates that even when precision in estimation is achieved, the 95% confidence interval can include values which are less than half and more than double the true effect magnitude (i.e., < 0.1 and > 0.4). While the four outcomes shown in Figure 2.3 have a low probability of occurrence, the probability of obtaining 95% a confidence interval for the effect which contains the null value is above 15% in this example. Crucially, the probability of being unable to make directional conclusions about population-level effects, or obtaining intervals which contain grossly exaggerated effect values, is not explicitly quantified by a precision analysis.

The consequences of selecting a target width on the capacity to draw directional conclusions about a population-level effect may not be immediately obvious to the researcher when using a precision-based approach to identify the required sample size. To guard against the possibility of selecting a level of precision which precludes directional conclusions, the researcher may select a conservative value for *w* to ensure confidence in obtaining narrow intervals which are located close to the true effect. However, high precision requires large sample sizes. For example, considering the example above, if *w* is decreased from 0.3 to 0.1, the sample size required to achieve this level of precision in estimation on 80% of occasions increases from 36 to 268. While a highly stringent level of precision can provide confidence in obtaining accurate estimates for an effect which do not include the null, factors such as insufficient funds, risks of participation or a limited pool of participants can render this level of precision unachievable. Further, selecting an arbitrarily conservative *w* to permit directional conclusions may result in a sample size which is larger than that required if, instead, the locations of the effect intervals are considered alongside the null value. It is only in the context of the magnitude of the values contained within and excluded from the effect interval that a researcher can consider whether the target level of precision is either too low or beyond what they require.

2.3.1 An Alternative Approach to Sample Size Selection

Considering the shortcomings of estimation goals which are focused solely on either rejecting the null hypothesis or on achieving precision in estimation, an alternative goal of estimation is adopted within this thesis. Specifically, the aim here is to obtain accurate and precise estimates for the effects of interest which permit directional inferences about population-level differences. In the context of the current research and the chosen approach to measurement and analysis, this goal corresponds to obtaining an accurate and precise estimate for the regression slope coefficient which indicates whether or not metacomprehension judgements are predictive of understanding on health-related texts at the population-level. Given this goal of estimation, a method is required to quantify the a priori probability of achieving this goal which formally considers the magnitudes of effect estimates alongside the widths of the effect intervals.

Previously, Jiroutek et al. (2003) have proposed a method of sample size estimation which considers the probability of achieving a goal of estimation which is similar to that adopted in this thesis. Jiroutek et al. (2003) identify three criteria which define the successful estimation of an effect: width, validity, and rejection. The width criterion is consistent with that used in precision analysis to satisfy a target level of uncertainty in effect estimation. The validity criterion is met if the population effect is located within the effect interval, while rejection occurs if the effect interval excludes a specified null value. While this approach considers both accuracy, precision and exclusion of the null value, it is arguably more useful to know the probability of obtaining an effect estimate within a fixed magnitude from the true effect, rather than the probability of obtaining values which fall somewhere inside effect intervals of a specified width. For example, for larger effects, these three criteria would not preclude grossly over or underestimated effects, if the width criterion is not sufficiently conservative. Similar to precision analysis, therefore, Jiroutek et al.'s (2003) approach is limited by the lack of a formal consideration of the magnitude of the estimated effect, relative to the true population-level effect, alongside precision.

Beyond approaches to sample size estimation, an interest in both magnitude and precision can be found in what is sometimes referred to as parameter recovery analysis. Parameter recovery analysis is typically undertaken in the context of comparing the performance of competing theoretical or analytic models (e.g., Bürkner & Charpentier, 2020). Typically, alongside measures of precision, measures of the magnitude of estimates are calculated to evaluate accuracy in parameter recover analysis. For example, measures of root-mean square error provide an estimate of the average level of discrepancy between the parameter estimates and true parameter value, whereas measures of bias capture the overall tendency to over or underestimate the true parameter value (Walther & Moore, 2005). Unfortunately, these metrics of the accuracy of an estimate only provide insight into how a parameter is recovered on average across hypothetical studies. In the context of designing a study to obtain accurate and precise estimates of an effect, it would be much more useful to know the probability of observing effect magnitudes which are located close to the population-level effect.

Given that a method which directly considers the sampling distributions of both the point estimates and interval widths in determining sample size has not previously been described, building on the approaches advocated by Jiroutek et al. (2003) and Maxwell et al. (2008), a novel approach is presented here. Addressing concerns with statistical power and precision analysis, this approach plans for the capacity to obtain high quality estimates of population-level effects, whilst balancing i) the desire to make directional conclusions and ii) demands on sample size requirements. In the remainder of this chapter, this approach to determining sample size is presented. Firstly, three steps are described which operationalise the probability of achieving both accuracy and precision in estimation. Following this, to demonstrate the approach, a closed-form solution is presented for identifying the required sample size to estimate the difference between two groups in the mean of a variable of interest. Lastly, simulated observations for the same example scenario are presented to aid in further illustrating the approach.

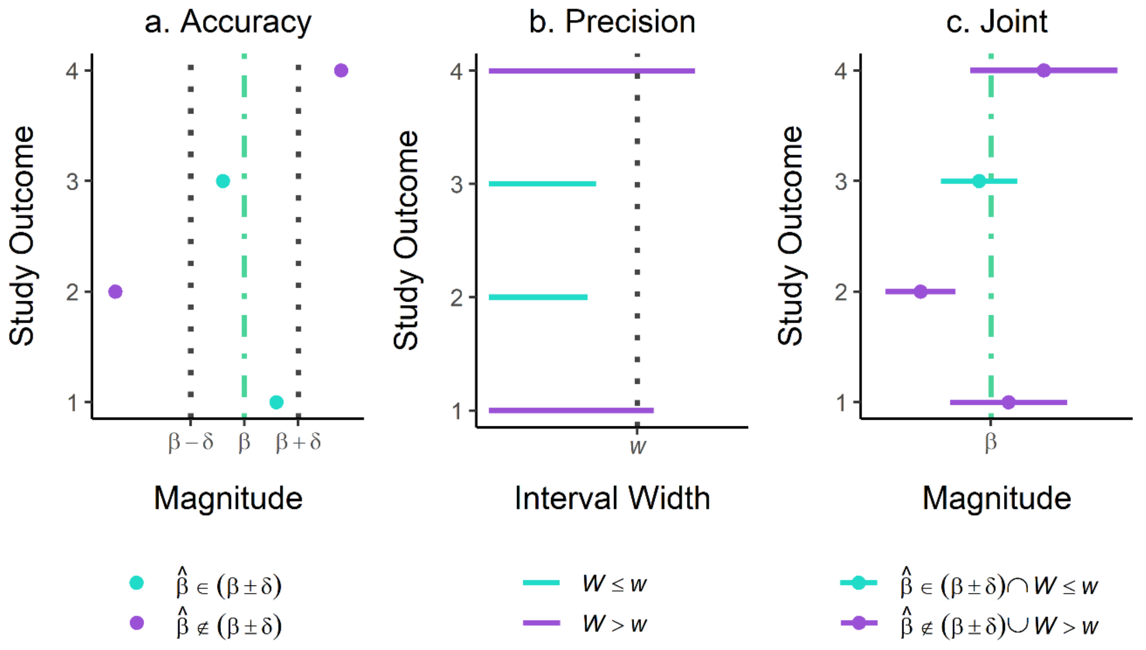
Accuracy and Precision in Estimation. Firstly, a target level of accuracy is defined by a fixed difference δ from the magnitude of the population-level effect of interest β. This difference δ is used to construct an accuracy interval around the effect parameter β, to evaluate the accuracy of the estimate of the effect $\hat{\beta}$. Satisfactory accuracy is achieved when the estimate $\hat{\beta}$ falls within the region around β defined by δ:

$$\beta - \delta \geq \hat{\beta} \leq \beta + \delta \,. \qquad (1)$$

This accuracy criterion is illustrated in Figure 2.4a, showing that effect estimates which fall within the accuracy interval defined by $\beta \pm \delta$ are accepted as sufficiently accurate. For effect estimates which are not an element of the set of values defined within the accuracy interval, these effect estimates are rejected.

Figure 2.4

Illustration of Accuracy and Precision Acceptance Criteria and their Joint Application in Determining the Success of Effect Estimation



Note. Effect estimation is considered a success (shown in teal) when the criteria of accuracy (a.), precision (b.) and their joint application (c.) are met. Otherwise, effect estimation is considered unsuccessful (shown in purple). β = population-level effect, $\hat{\beta}$ = point estimate of effect, δ = difference to define accuracy interval, w = target interval width, W = observed interval width.

Secondly, a target level of precision is defined according to a fixed width w for the interval for the effect. Consistent with Kelley et al. (2003) and Maxwell et al. (2008), the selected width is used to assess the precision of the estimated effect, by comparing w to the observed width W of the effect interval for β. The absolute difference between the upper interval value β_{ub} and the lower interval value β_{lb} for the effect provides the measure of W:

$$|\beta_{\text{ub}} - \beta_{\text{lb}}| = W\,.$$

Precision is determined according to whether W equal to or less than w. Satisfactory precision is achieved when:

$$W \leq w\,. \qquad (2)$$

This precision criterion is illustrated in Figure 2.4b, demonstrating that observed widths which do not exceed the target width w are accepted as sufficiently precise. In contrast, observed widths which exceed w are rejected.

Lastly, the accuracy and precision criteria are combined, to jointly evaluate the probability of estimating the effect of interest with both accuracy and the precision:

$$\mathrm{P}(\hat{\beta} \in (\beta \pm \delta) \cap W \leq w)\,. \tag{3}$$

The joint acceptance criteria are illustrated in Figure 2.4c. Successfully achieving the goal of estimation occurs when the effect estimate $\hat{\beta}$ is simultaneously within $\pm\,\delta$ from β and the observed width W of the interval for the effect is less than or equal to w. If either the accuracy or precision criteria, or both, are not met, rejection occurs.

Example: Difference Between the Means of Two Groups. To demonstrate this novel approach to quantifying the probability of estimating an effect with accuracy and precision, a simple experimental design and analysis is described here. In this example, assume that the researcher is interested in comparing the average value of a variable of interest between two groups of participants. As described in section 2.2, we could consider the two groups correspond to a control group and a group which receives a situation-model based intervention aimed at improving metacomprehension accuracy (Prinz et al., 2020b). This scenario demonstrates how the framework of jointly estimating with accuracy and precision can be used to derive closed-form solutions to determine the required sample size.

For simplicity, in this example, assume that each participant is placed in either the control or treatment group and provides one observation of the variable of interest. Further, assume that responses are independent and identically distributed within each group. Let Y_{ij} denote the response variable of interest, where i = 1,2 indexes the group (1 = control, 2 = intervention) and j = 1,…,n_i indexes individuals, and assume that the responses are normally distributed with mean μ_i and common (shared) standard deviation σ:

$$Y_{ij} \sim N(\mu_i, \sigma).$$

Further assume, without loss of generality, that control and intervention groups have equal numbers of participants (with $n_1 = n_2 = n$) and hence we can express the group-specific observations as response vectors each of length n:

$$Y_1 = (y_{11}, y_{12}, \dots, y_{1n})^T,$$

$$Y_2 = (y_{21}, y_{22}, \dots, y_{2n})^T.$$

For each group, the sample mean then provides an estimate of the population mean and is given by:

$$\hat{\mu}_1 = \frac{y_{11} + \cdots + y_{1n}}{n},$$

$$\hat{\mu}_2 = \frac{y_{12} + \ldots + y_{2n}}{n}.$$

The sample variance for each group is then calculated as:

$$\hat{\sigma}_1^2 = \frac{\sum_{j=1}^{n} \left(y_{1j} - \hat{\mu}_1\right)^2}{n-1},$$

$$\hat{\sigma}_2^2 = \frac{\sum_{j=1}^{n} \left(y_{2j} - \hat{\mu}_2\right)^2}{n-1},$$

and the pooled estimate of the underlying population variance, σ^2, is given by

$$\hat{\sigma}_P{}^2 = \frac{(n-1)\hat{\sigma}_1^2 + (n-1)\,\hat{\sigma}_2^2}{2n-2}. \tag{4}$$

A $(1 - \alpha)\%$ confidence interval for the true difference in group means is then given by:

$$\left[(\hat{\mu}_1 - \hat{\mu}_2) - t_{2n-2,1-\frac{\alpha}{2}}\,\hat{\sigma}_P, . (\hat{\mu}_1 - \hat{\mu}_2) + t_{2n-2,1-\frac{\alpha}{2}}\,\hat{\sigma}_P\right],$$

where $t_{2n-2,1-\frac{\alpha}{2}}$ is the critical value of the t-distribution with $2n - 2$ degrees of freedom and a specified significance level α.

Given response vectors Y_1 and Y_2, group membership can be characterised using a binary indicator variable, x_{ij}, where x_{1j} = 0 (control group) and x_{2j} = 1 (intervention group). Responses y_{ij} can then be expressed in the form of a general linear model:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}, \tag{5}$$

where β_0 and β_1 are (unknown) parameters of interest and the error terms ε_{ij} are independent and identically distributed normal random variables with mean zero and standard deviation σ:

$$\varepsilon_{ij} \sim N(0, \sigma).$$

Note that, given the group indicator x_{ij}, the linear model parameters in (5) map directly on to the population means:

$$\beta_0 = \mu_1 ,$$

$$\beta_1 = \mu_2 - \mu_1,$$

and correspondingly the model estimates of the parameters β_0 and β_1 map on to the sample means:

$$\hat{\beta}_0 = \hat{\mu}_1,$$

$$\hat{\beta}_1 = \hat{\mu}_2 - \hat{\mu}_1.$$

Given the accuracy acceptance criterion defined in (1), in the context of the present example, the probability of satisfying (1) can be calculated based on the sampling distribution of $\hat{\beta}_1$, where:

$$\hat{\beta}_1 \sim N\left(\beta_1, \sqrt{\frac{2\sigma^2}{n}}\right).$$

Standardising, the probability that $\hat{\beta}_1$ falls within the interval defined by $\beta_1 \pm \delta$, centered at the true difference, can hence be evaluated:

$$\mathrm{P}(\hat{\beta}_1 \epsilon (\beta_1 \pm \delta) \mid \delta, \sigma, n) = \int_{-\infty}^{z} \left(\frac{\delta}{\sqrt{\frac{2\sigma^2}{n}}}\right) dz - \int_{-\infty}^{z} \left(\frac{-\delta}{\sqrt{\frac{2\sigma^2}{n}}}\right) dz$$

$$= \Phi\left(\frac{\delta}{\sqrt{\frac{2\sigma^2}{n}}}\right) - \Phi\left(\frac{-\delta}{\sqrt{\frac{2\sigma^2}{n}}}\right),$$

where Φ (.) is the cumulative distribution function (CDF) of the Standard Normal distribution.

With respect to precision, the observed width W is calculated as the absolute difference between the upper and lower values of the interval for the effect. In the context of the present example, the upper and lower bounds are calculated based on the estimate $\hat{\beta}_1$, the standard error SE of the estimate and the critical value of $t_{2n-2,1-\frac{\alpha}{2}}$:

$$\beta_{1ub} = \hat{\beta}_1 + t_{2n-2,1-\frac{\alpha}{2}} SE,$$

$$\beta_{1lb} = \hat{\beta}_1 - t_{2n-2,1-\frac{\alpha}{2}} SE,$$

where

$$SE = \sqrt{\hat{\sigma}_P{}^2}\sqrt{2/n}\,.$$

Assuming a $(1 - \alpha)\%$ confidence interval is calculated for β_1, the observed width W can then be calculated as the difference between the upper and lower bounds:

$$W = \left(\hat{\beta}_1 + t_{2n-2,1-\frac{\alpha}{2}}\left(\sqrt{\hat{\sigma}_P{}^2}\sqrt{2/n}\right)\right) - \left(\hat{\beta}_1 - t_{2n-2,1-\frac{\alpha}{2}}\left(\sqrt{\hat{\sigma}_P{}^2}\sqrt{2/n}\right)\right) \qquad (6)$$

and precision is achieved when $W \leq w$.

Utilising (4) and (6), the sum of the sample variance estimates can be expressed:

$$\hat{\sigma}_1^2 + \hat{\sigma}_2^2 = 2\left(\frac{W}{2t_{2n-2,1-\frac{\alpha}{2}}\sqrt{2/n}}\right)^2$$

and hence precision will be achieved when the inequality:

$$\hat{\sigma}_1^2 + \hat{\sigma}_2^2 \leq 2\left(\frac{w}{2t_{2n-2,1-\frac{\alpha}{2}}\sqrt{2/n}}\right)^2$$

is satisfied. Specifically, only values for the summed sample variance which produce $(1-\alpha)$% confidence intervals for β_1 with width $\leq w$ satisfy this inequality. The probability of observing combined sample variances which satisfy this inequality can be calculated based on the sampling distribution of $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$.

For each sample from Y_1 and Y_2, given the common population variance σ^2, the variances of the samples $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ will be distributed:

$$\hat{\sigma}_1^2\frac{n-1}{\sigma^2} \sim {\chi^2}_{n-1},$$

$$\hat{\sigma}_2^2\frac{n-1}{\sigma^2} \sim {\chi^2}_{n-1}.$$

Given that the sum of two Chi-square random variables are also Chi-square distributed, the combined sample variances $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ will also be Chi-square distributed:

$${\chi^2}_k + {\chi^2}_l \sim {\chi^2}_{k+l},$$

$$(\hat{\sigma}_1^2 + \hat{\sigma}_2^2)\frac{n-1}{\sigma^2} \sim {\chi^2}_{2n-2}.$$

The expression for combined sample variances, in terms of the observed $(1-\alpha)$% confidence interval width, can then be substituted to provide the distribution with respect to the observed width:

$$\hat{\sigma}_1^2 + \hat{\sigma}_2^2 = 2\left(\frac{W}{2t_{2n-2,1-\frac{\alpha}{2}}\sqrt{2/n}}\right)^2,$$

$$(\hat{\sigma}_1^2 + \hat{\sigma}_2^2)\frac{n-1}{\sigma^2} \sim {\chi^2}_{2n-2}.$$

Therefore,

$$\left(\frac{W}{2t_{2n-2,1-\frac{a}{2}}\sqrt{2/n}}\right)^2\left(\frac{2n-2}{\sigma^2}\right) \sim {\chi^2}_{2n-2}\,.$$

In the context of the present example, given the precision acceptance criterion defined in (2), the probability of observing a $(1-\ \alpha)\%$ confidence interval for β_1 of width ≤ w can then be calculated as the area under the Chi-square distribution with degrees of freedom $2n-2$, given by:

$$\mathrm{P}(W \leq w \mid w, \sigma, n) = \int_0^u \left(\frac{w}{2t_{2n-2,1-\frac{a}{2}}\sqrt{2/n}}\right)^2\left(\frac{2n-2}{\sigma^2}\right) du\,.$$

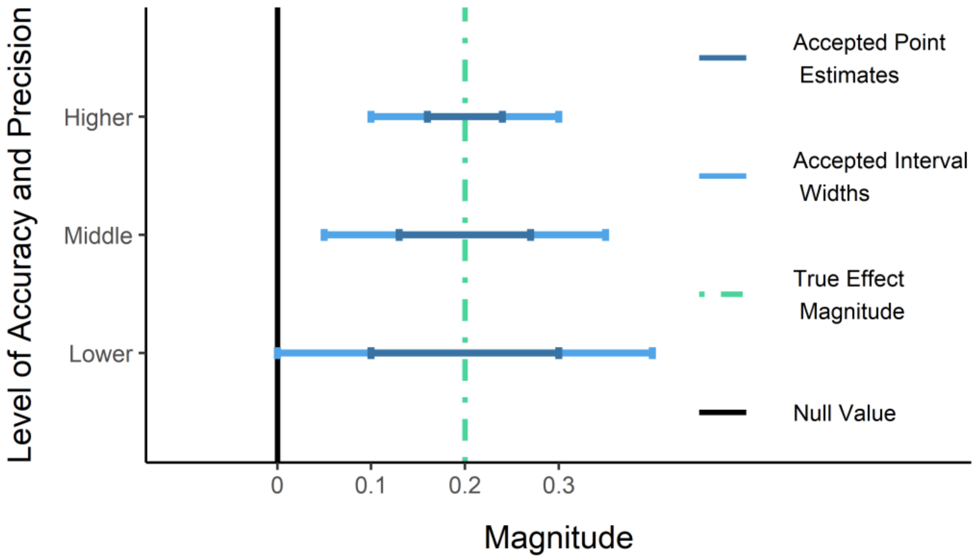
The probability of jointly achieving both accuracy and precision in estimation can then be calculated as the product of the probabilities of satisfying the accuracy acceptance criterion and the precision acceptance criterion:

$$\mathrm{P}(\hat{\beta}_1 \,\epsilon\, (\beta_1 \pm\ \delta\ )\ \cap\ W \leq w \mid \delta, w, \sigma, n) = \mathrm{P}(\hat{\beta}_1 \,\epsilon\, (\beta_1 \pm\ \delta\ )) \times \mathrm{P}(W \leq w).$$

Having defined the closed-form solution to estimate a difference in the average value of a variable of interest between two groups, with accuracy and precision, the required sample size can readily be calculated. For this example, values for δ, w, and σ must be selected to calculate n. Further, to consider the capacity to make directional conclusions, a value for the population-level effect β_1 should also be selected. Adopting the values previously selected in section 2.2, σ = 0.2 and β_1 = 0.2. Three pairs of values for δ and w are selected in the current example to reflect relatively **lower** (δ = 0.1, w = 0.2), **middle** (δ = 0.07, w = 0.16) and **higher** (δ = 0.04, w = 0.12) levels of accuracy and precision, to demonstrate the implications on parameter estimation and n. Figure 2.5 illustrates how these pairs of values for δ and w define the range of effect estimates which will be accepted relative to the population-level effect value. Given the **lower** level of accuracy and precision, negative values are excluded from the effect interval, but the null value is not. In contrast, given the **higher** level of accuracy and precision, point estimates less than half of the population-level effect magnitude are considered inaccurate. The **middle** level of accuracy and precision captures a mid-point between these criteria.

Figure 2.5

Accepted Magnitudes for the Effect of Interest Given Varying Accuracy and Precision Levels



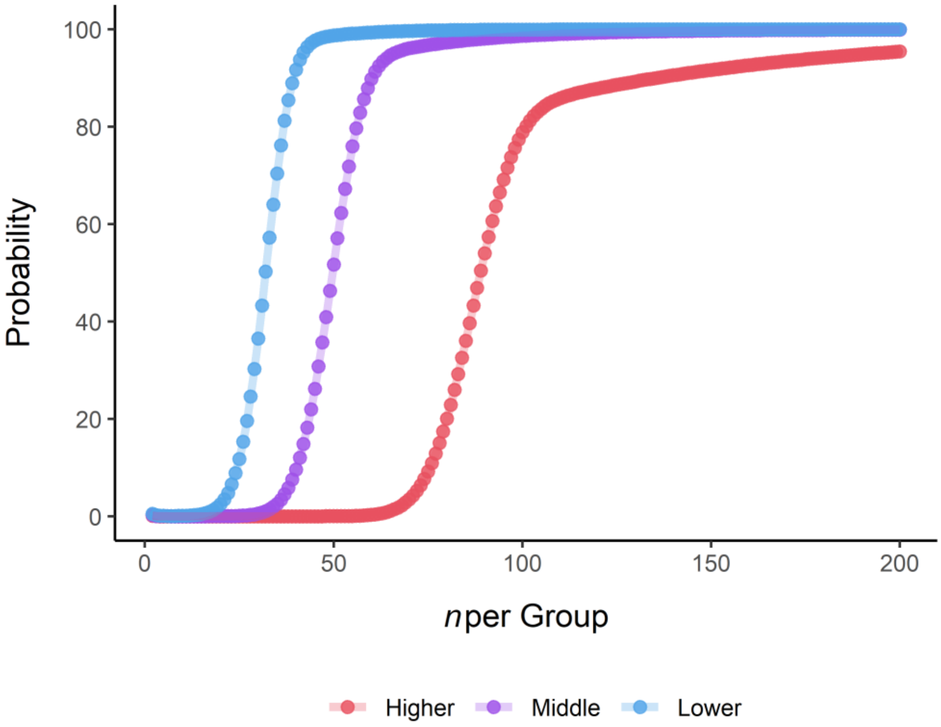
Note. The range of values depicted as ‘accepted point estimates’ illustrates the interval captured by the population-level effect $\pm\,\delta$. The range of values depicted as ‘accepted interval widths’ illustrates the interval of values captured by the maximally accepted interval width ($W = w$) given a point estimate for the effect located at the upper and lower bounds of the accuracy interval.

Correspondingly, the probability of jointly satisfying the accuracy and precision criteria in this example can be calculated across sample sizes for each of the three selected levels of accuracy and precision. Figure 2.6 shows the power curves for each of the selected values of δ and w, given samples sizes of four to 200 participants per group. Assuming an 80% target probability of achieving the goal of estimation in quantifying the difference in the average value of a variable of interest between two groups, a total sample size of 74, 114, and

202 participants are required, given the **lower**, **middle** and **higher** levels of accuracy and precision selected respectively.

Figure 2.6

Probability of Achieving Varying Levels of Accuracy and Precision in Estimation



Simulating Observations to Determine Sample Size. The most efficient approach to quantify the probability of achieving the goal of estimation is to use a formula based on the probability distributions of the relevant parameters. Often, however, due to variation and complexity in the experimental design and planned analysis, applicable formula may not be available or, where one has been derived, it may be challenging to locate within the literature (Arnold et al., 2011; Johnson et al., 2015; Landau & Stahl, 2013; McConnell & Vera-Hernández, 2015). In such situations, the alternative approach is to use simulations to estimate of the probability of achieving the estimation goal. Given the multilevel design and Bayesian estimation framework which will be adopted within this thesis, simulation is required to provide an estimate of the required sample size. To demonstrate how simulation can be used to estimate the probability of estimating with accuracy and precision, the same example concerning the estimation of a difference in the average value of a variable of interest between two groups will be used. However, this will be conducted within a Bayesian estimation framework to further illustrate how this approach to sample size determination can be applied.

To simulate data, a model which describes the process by which observations are generated must first be defined. After selecting an appropriate model, parameter values must be selected in order to simulate observations. Following this, for each n in a set of candidate sample sizes, a vector of responses are simulated. The planned statistical analysis of the empirical data is then applied to the simulated data and estimates of the effect of interest are obtained. The estimates are then evaluated against the acceptance criteria used to determine whether the simulated study resulted in an estimation success or a failure. Multiple simulations of each candidate n is conducted, typically 1000 times, reflecting an attempt to balance computational time and variability across simulations. The number of successful estimation outcomes as a proportion of the total simulations conducted provides an estimate of the probability of achieving the goal of estimation.

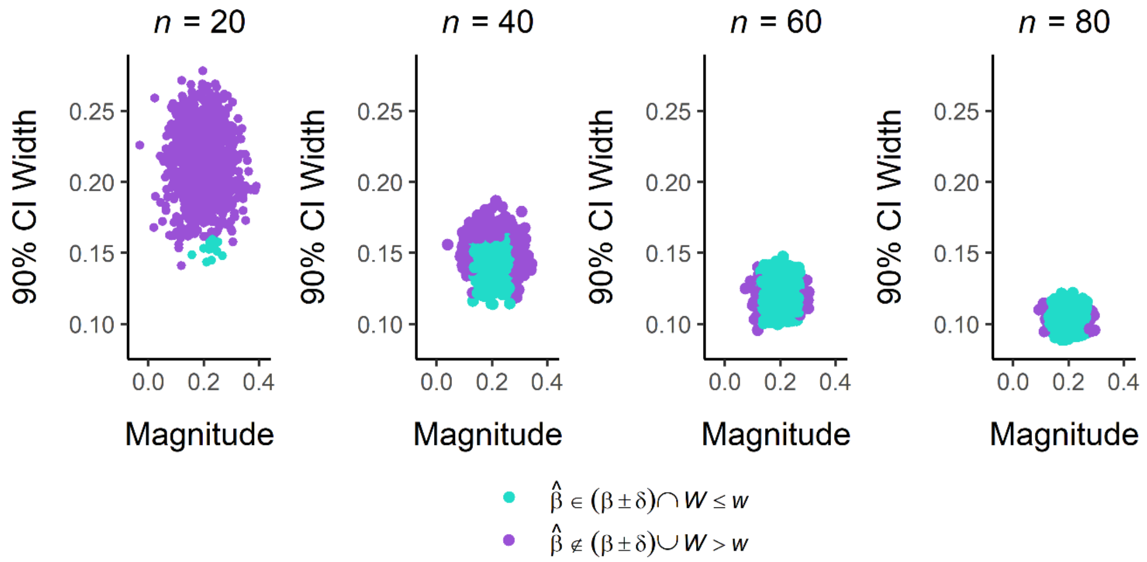
In the present example, the response variable of interest is again assumed to be normally distributed, with equal variance for both the control and the intervention group. Given this, a linear model formulation, described in (5), serves as the generative model of the response variable. Consistent with the values selected previously: x_{1j} = 0 (control group) and x_{2j} = 1 (intervention group), $\beta_1 = 0.2$, and $\sigma = 0.2$. In addition, a value for β_0 is selected to fully parameterise the model: $\beta_0 = 0.3$. Given these values, observations for the control group were simulated as $N(\mu = 0.3, \sigma = 0.2)$ and intervention group observations were simulated as $N(\mu = 0.5, \sigma = 0.2)$. Candidate sample sizes per group of n = (20, 40, 60, 80) were each simulated 1000 times.

A Bayesian linear regression model was then fit to each simulated dataset, using weakly informative prior distributions for the model parameters. Weakly informative prior distributions refer to probability distributions which are relatively flat across the range of values which a parameter can plausibly be expected to take. In this example, for the population-level parameters (β_0, β_1), normal distributions with mean 0 and standard deviation 10 were specified and a half-student-t distribution, with the values of the degrees of freedom, location and scale parameters set to 3, 0 and 10, was specified for the standard deviation. From each model, the mean of the posterior distribution was used to evaluate the accuracy of the estimated effect and the width of the 90% credible interval was used to evaluate the precision of the estimate.

Based on this simulation, given δ = 0.07, w = 0.16, the probability of simultaneously achieving accuracy and precision was estimated to be 1.3%, 75%, 95% and 97% for sample sizes per group of n = (20, 40, 60, 80), respectively. To further illustrate how the accuracy and precision in estimation approach relates to the quality of the effect estimates accepted and rejected, the simulated estimated effects are displayed in Figure 2.7, coloured according to the joint accuracy and precision acceptance criteria. In this figure, each point corresponds to the result of one simulation, plotted according to the magnitude of the posterior mean and the width of the 90% credible interval. For readability, the magnitudes of values contained within the intervals for the effect are not directly plotted against the horizontal axis. As can be seen in Figure 2.7, as n increases, the estimated magnitude of the posterior means move closer to the population-level effect (0.2) and the width of the effect interval reduces, reflecting a reduction in sampling variability. At $n = 40$, the impact both of the accuracy and precision criteria are evident, as many intervals are estimated with satisfactory width, yet are not centred sufficiently close to the population-level effect, whilst several intervals are estimated with insufficient precision despite being located close to the population-level effect.

Figure 2.7

Simulated Effect Estimates and Interval Widths Given Varying Sample Sizes



Note. *n* = number of participants per group. CI = credible interval.

2.4 Summary

This chapter provides a clear and evidenced basis for the selection of the measurement of metacomprehension accuracy, the analysis of differences in metacomprehension accuracy, and the identification of the required sample size to address the research questions within this thesis. As described in section 2.1.1, metacomprehension accuracy will be operationalised as the population-level slope coefficient of a regression model. Further, as detailed in section 2.2.1, the regression model fit to the empirical data will be a Bayesian logistic multilevel model. Lastly, to adequately plan for the capacity to address the research questions, the novel approach presented in section 2.3.1 will be used to identify the sample size required to simultaneously estimate metacomprehension accuracy with accuracy and precision, whilst providing the capacity to draw directional conclusions. Together, the selected approaches to measurement, analysis and sample size determination allow for the capacity to obtain empirical evidence which permits clear inferences to be drawn concerning the predictive validity of reader panellist's judgements of comprehension and the utility of metacomprehension judgements in the review of health-related texts.

3. Study 1

In this chapter, the motivation for the first study (Study 1) is briefly discussed, leading to the identification of two research questions to be addressed. In the method section, two pilot studies are first described (Pilot 1 and 2), which inform the task specification and sample size requirements of Study 1. The identification of the required sample size is considered in a design analysis within each pilot. Following this, the participants, materials and procedure of Study 1 are described. The presentation of results is separated into a preliminary data inspection, the main planned analysis, and an analysis of the sensitivity of the main findings to alternative analytical choices. Lastly, the findings of Study 1 are discussed.

3.1 Introduction

To consider whether reader panellists judgements are likely to be a valid measure of the comprehensibility of health information, Study 1 was conducted to quantify the predictive relationship between perceived and assessed comprehension. Two aspects of the relationship were of primary interest: i) the overall predictive relationship across individuals and ii) the between-individual variability in this relationship, if any. In addition to this, information was collected to explore previous suggestions that differences between individuals in comprehension-related abilities underlie variation in metacomprehension accuracy. Specifically, Study 1 was designed to explore whether the relationship between perceived and assessed comprehension interacts with differences in individual's reading ability and background knowledge. These variables were selected due to i) the role of domain-specific knowledge and general comprehension-based skills, such as inferencing and reading strategies, in multiple models of reading comprehension (Ahmed et al., 2016, Cromley & Azevedo, 2007; Gough & Tunmer, 1986; Perfetti, 1999; Perfetti & Stafura, 2014; Tzeng et al., 2005; van den Broek & Helder, 2017) and ii) previous conflicting accounts of their influence on metacomprehension accuracy (Glenberg & Epstein, 1987; Griffin, Wiley & Thiede, 2019; Griffin et al., 2009; Jee et al., 2006; Lin et al., 2002; Löffler et al., 2016; Maki et al., 2005).

3.1.1 Research Aims of Study 1

Given the above, Study 1 was designed to address two research questions:

RQ1: Are judgements of comprehension predictive of assessed comprehension on health-related texts?

RQ2: How are individual differences in reading ability and background knowledge related to variation in the relationship between perceived and assessed comprehension?

3.2 Method

Ethical approval for this study was gained (for piloting and data collection) in December 2018. Piloting commenced in January 2019 and data collection began in February 2020. A detailed preregistration for this study, including materials and analysis plan, was uploaded to an OSF repository in February 2020 (https://osf.io/b8fex/).

Prior to full data collection, two pilot studies were conducted. Pilot 1 aimed to inform on several design concerns; Pilot 2 explored issues arising from the initial pilot. Full details of Pilots 1 and 2 are provided in sections 3.2.1 and 3.2.2, respectively

3.2.1 Pilot 1

Three main design concerns were explored in Pilot 1: i) the effect of open-ended and multiple-choice question response formats on task performance, ii) the impact of text-present and text-absent conditions on task performance, and iii) the framing of items used to elicit judgements of comprehension. In addition, this pilot study provided evidence to inform an assessment of the number (sample size) required of participants and texts to ensure a high probability of estimating the effects of interest with accuracy and precision.

Participants. A sample of 20 UK nationals was recruited using the online platform Prolific (Prolific.co). Age ranged from 20 to 56, with mean *M* = 28.85 and standard deviation *SD* = 8.33). Fifteen participants identified as female and 5 as male. Participation was restricted to those whose highest level of qualification did not exceed A-Level standard. Participants received £8.00 compensation for taking part.

Materials.

Health Texts. Health information leaflets, concerning health conditions and medical treatments, were identified online via opportunity sampling from various NHS trust's patient information websites. To limit topic familiarity, texts were selected which were considered less likely to be well known. Potentially embarrassing topics were not included. From this sample, three texts were chosen for the pilot: femoral hernia, basal cell carcinoma and percutaneous liver biopsy. The structure of headings and font formatting (i.e., emboldened text) was preserved. Contact information, images, and text-references to images were removed. Descriptive information for each text is provided in Table 3.1 and the full texts are provided in Appendix A.

Table 3.1

Descriptive Information for Health Texts in Pilot 1

Topic	Words	Flesch Score	Flesch-Kincaid
Femoral hernia	1285	39.7	11.3
Basal cell carcinoma	1033	66.6	7.2
Percutaneous liver biopsy	1150	62.8	8.7

Note. Flesch score refers to the Flesch Reading Ease score. Flesch-Kincaid refers to the Flesch-Kincaid Grade level. Calculated using the online tool Coh-Metrix 3.0 (Graesser, et al., 2004).

Five comprehension questions were developed for each text which aimed to test the situation model level of understanding (Wiley et al., 2005). Questions were formed by identifying information not explicitly stated in the text, but which could be inferred. The correct answer to the question required an inference, connecting pieces of information provided in the text. To reduce variability in question complexity, attempts were made to limit the background knowledge necessary to establish the inference. For example, for the first question on the basal cell carcinoma text (‘Why might basal cell carcinomas be less likely to occur on the feet?’), the required pieces of information from the text were the propositions ‘BCCs arise due to too much sun exposure’ and ‘They are more likely to occur on body areas that catch the sun’. The required inference that linked the propositions was that basal cell carcinomas grow on body areas that are exposed to too much sun. The additional minimal background knowledge required to answer the question concerned the feet typically being covered up. Together, these sources of information were considered sufficient to infer the correct answer: ‘The feet are more likely to be covered up and less exposed to the sun, so basal cell carcinomas are less likely to occur there’.

For the multiple-choice question format, three types of distractors were created: literal, misconception, and near-miss inference. Literal distractors consisted of information explicitly stated in the text (i.e., the information that the correct inference is based upon). Misconceptions were common incorrect responses made in the open-ended question format, though these could be factually correct but did not answer the question. Near-miss inferences captured an inference which did not answer the question. The multiple-choice questions and response options are provided in Appendix B.

Perceived Comprehension. To consider variability in responses to judgement prompts (Pilegard & Mayer, 2015), four prompts were used to elicit metacomprehension judgements in the first pilot study: ‘Overall, how well do you understand the text?’, ‘How much of the text do you feel you understand?’, ‘How well do you think someone else would understand the text?’, and ‘How easy was it to understand the text?’. A 5-point Likert scale, presented as labelled responses, was designed for each prompt. The following labels were used for the response options, given the respective prompt: 'not well at all/none at all/extremely difficult', 'slightly well/a little/somewhat difficult', 'moderately well/a moderate amount/neither easy nor difficult', 'very well/a lot/somewhat easy', and 'extremely well/all of it/extremely easy'.

Qualitative Reading Inventory (QRI). To provide a measure of reading ability, materials from the QRI sixth edition (Leslie & Caldwell, 2017) were used. The QRI is an informal reading inventory, containing assessment materials for 1st to 6th grade, upper-middle school and high school, which can be flexibly used as a diagnostic tool for reading instruction or to determine standardised reading level. In identifying the appropriate elements to select from the QRI, the reading ability in the UK adult population was considered. Based on the 2011 UK census data, the Office for National Statistics (2013) report the following percentages of qualifications within the population: no qualifications (23%), level 1 (14%), level 2 (15%), level 3 (12%) and level 4 and above (27%). Similarly, an analysis of the Programme for the International Assessment of Adult Competencies (PIAAC) 2012 dataset, reported by the Organisation for Economic Co-operation and Development (Kuczera, Field, & Windisch, 2016), found that over 15% of the adult population in England had literacy skills below level 2. The modal highest level of qualification may, therefore, be between level 2 and level 3, equivalent to achieving between five or more GCSEs or two or more A-levels. Based on this information, the assessment texts in the QRI which corresponded to upper-middle school (described below) were selected.

The upper-middle school, level-diagnostic reading assessment materials in the QRI comprise three texts: a narrative text which tells the biographical story of a teacher and two expository (non-fiction) texts about American immigration and the life cycle of stars. The two expository texts were selected for use, given their greater similarity to the health-related texts. The American immigration and life cycle of stars texts are 839 and 802 words in length, respectively. Each text contains two passages, with ten open-ended comprehension questions to be administered after each passage is read. The comprehension questions on each passage are designed to measure explicit and implicit levels of understanding. A sum score was calculated to provide a measure of reading ability, given by the total performance across both passages, with a maximum score of 20 per text.

Given the intention to implement the QRI materials in an online testing environment, it was considered whether participant's motivation to provide open-ended responses may impact response accuracy. To explore this, multiple-choice response versions of the 20 comprehension questions per text were designed to provide an alternative measure of reading ability, permitting a comparison of performance between open and closed response formats. The questions from the QRI assessment materials were used for the multiple-choice questions. The correct response options constructed from the answers provided in the QRI manual and three incorrect response options (distractors) were created per question.

Distractors were created according to three categories, to improve the effectiveness of distractors and the validity of correct responses: literal, misconception and near-miss inference (Downing & Haladyna, 1997; Graesser et al., 2009). Literal distractors reflected a surface-level understanding of the text with no evidence of an inference. Misconception distractors tapped a common misconception or misinterpretation given by participants in the open-ended response format of the task. Near-miss inference distractors captured an inference which did not fully answer the question, such as a plausible inference but which was not correct in the context. This format of distractor categories was adapted for explicit QRI questions. For these questions, for which the correct answer was literal, literal distractors were replaced with near-miss literal distractors (see Appendix C for examples). In addition, some explicit questions were not amenable to misconception or near-miss inference distractors. In these instances, incorrect response options with high lexical and semantic overlap with the text and correct response were used. The full list of the distractors is provided in Appendix C. Performance was calculated as the total number of correct responses selected across both passages, with a maximum score of 20 per text.

Health Literacy Vocabulary Assessment (HLVA). To provide a measure of health-related background knowledge, the HLVA (Ratajczak, 2020) was chosen. The HLVA is designed to provide a measure of health-related literacy in individuals with English as their first or second language. The HLVA comprises a list of 22 health-related words, for which participants provide a verbal definition after hearing each word. Definitions for each word are marked against criteria derived from dictionary definitions. Each word is scored as either incorrect (0), partially correct (1), or fully correct (2). A total score is calculated as the summation of scores across items. This knowledge-based test of topic-specific vocabulary was considered likely to discriminate between varying levels of expertise without the potential vulnerabilities of self-report measures or possible ceiling effects of alternative measures of health-related understanding (Davis et al., 2006; Ratajczak, 2020). Notwithstanding this, an analysis of HLVA response data, acquired during an unpublished MSc dissertation project (Chadwick, 2018), suggested that a number of items lacked discriminative power and partial scoring, compared to binary scoring, did not offer additional sensitivity in a sample of 50 fluent, adult, English readers participants. Therefore, 6 items were removed and full-scoring was applied, yielding a maximum sum score of 16. The 16 HLVA items administered are provided in Appendix D.

To examine how task performance might vary with response format, two versions of the HLVA were used: an open-ended and a multiple-choice version. The open-ended version prompted participants to define the item with no limit on response length. A response was scored as correct if the definition satisfied the full-scoring assessment criteria, a partial definition was awarded a score of 0. The multiple-choice version consisted of four response options. The correct answer reflected the full definition in the assessment criteria. For reasons discussed previously (Graesser et al., 2009), three distractor items were created according to three categories: semantic, misconception and insufficient. Semantic distractors reflected a lexical or semantic relative of the word to be defined. Misconception distractors drew upon a common misconception or misinterpretation which occurred in open-ended responses. Insufficient distractors represented answers which provided a vague level of detail, equivalent to a partial definition. The list of multiple-choice responses is provided with the items administered in Appendix D.

Procedure. Participation in the task commenced when participants responded to the online invitation on Prolific.co. All tasks were computer-based and presented via Qualtrics. Prior to beginning the tasks, participants were provided with the following information: “In this study, you’ll be asked to define 16 health-related words. Next, you’ll read and judge how easy 3 health-related texts are to understand. Texts will be about medical conditions and procedures, such as cancer and biopsy. After this, you’ll be asked a few more questions about the health-texts. Lastly, you’ll read some general texts and answer questions about them”. Participants were then prompted to provide consent and report their age and gender.

Participants were randomly assigned to one of four conditions: open-ended-text-present (OTP), open-ended-text-absent (OTA), multiple-choice-text-present (MTP) and multiple-choice-text-absent (MTA). In the text present conditions, the health texts and QRI passages were available when responding to comprehension questions, whereas these were not available in the text absent condition. The administration of the open-ended and multiple-choice forms of the HLVA was the same in the text present and text absent conditions.

All participants first completed the HLVA. In the open-ended HLVA, participants were instructed to write a short description, whereas in the multiple-choice HLVA, participants were asked to select the answer that seemed most appropriate. Following this, participants were provided with the following instructions: “You will now read 3 health-related texts. After you read each text, you will be asked to judge your understanding. Please answer honestly.”. Judgements of comprehension were elicited immediately following the reading of each text. Participants were then presented with the comprehension questions for each text (with either the text present or absent, in open or multiple-choice response format), in the same order texts were presented for reading. Lastly, participants read and completed the questions on the American immigration and life cycle of stars texts from the QRI (with either the text present or absent, in open or multiple-choice response format).

Results and Discussion. Pilot 1 was undertaken to facilitate the selection of a task design informed by empirical evidence and to aid in evaluating the required sample size likely to achieve accuracy and precision in estimation. Note that given the small sample size, differences between conditions should be treated as weakly indicative of true differences. Descriptive statistics for performance across tasks, by response format and text presence, are provided in Table 3.2.

Table 3.2

Mean Performance on Each Task by Condition, with Standard Deviations in Parentheses.

Condition	HLVA	QRI Text A	QRI Text B	Health-Text Comprehension
OTP	6.40 (4.88)	12.20 (3.49)	12.00 (3.08)	1.80 (1.08)
OTA	7.00 (1.87)	10.60 (1.14)	11.20 (4.43)	2.00 (1.20)
MTP	9.20 (1.64)	11.40 (1.34)	14.80 (1.64)	1.40 (1.06)
MTA	7.60 (3.13)	9.40 (3.78)	12.40 (2.88)	1.53 (0.83)

Note. OTP = open-ended-text-present, OTA = open-ended-text-absent, MTP = multiple-choice-text-present, MTA = multiple-choice-text-absent. HLVA = Health Literacy Vocabulary Assessment, QRI Text A = Qualitative Reading Inventory – American immigration, QRI Text B = Qualitative Reading Inventory – life cycle of stars. Possible range of scores for tasks, left to right: 0-16, 0-20, 0-20, 0-5.

Average performance on the QRI tended to be slightly lower on the American immigration text (Table 3.2; QRI Text A) than the life cycle of stars text (Table 3.2; QRI Text B). However, the 95% confidence intervals for mean performance overlapped across text presence and response format conditions. Mean performance in text absent conditions, regardless of response format, was lower than in text present conditions. Differences in the average performance levels tentatively suggest that the multiple-choice version may introduce greater dissimilarities in comprehension performance between these texts. Further, the American immigration text may be more difficult than the life cycle of stars text.

With respect to the HLVA, the administration of the open-ended and multiple-choice forms of the HLVA was the same in the text present (OTP and OTA) and text absent conditions (MTP and MTA). Average performance on the HLVA was slightly higher on the multiple-choice compared to the open-ended response format, though the 95% confidence intervals for these means overlapped. Examining open-ended responses to the HLVA indicated that a number of participants provided responses which matched verbatim definitions from online searches. These copied definitions were typically incorrect.

In the health-text comprehension task, average performance was marginally higher on the open-ended version, regardless of text presence condition. However, 95% confidence intervals overlapped across conditions. Average performance in both response formats was slightly higher when the text was absent. Examining the errors made in the multiple-choice format indicated that participants typically selected the literal distractor.

The alignment between responses to the prompts used to elicit judgements of perceived comprehension were evaluated across all pilot conditions. Given that the judgements may not constitute interval-level data or be linearly related, Kendall's Tau-*b* (Kendall, 1955) was used to quantify the extent to which higher values on one prompt are associated with higher values on another prompt (see Table 3.3). Judgement prompts 1, 3 and 4 appeared to align most strongly, while prompt 2 showed lower associations. This suggests that judgements requiring participants to quantify the amount of text they understood was, to some extent, influenced by sources of variance not shared with the three other prompts. This pattern of association may occur due to differences in response wordings between prompts.

Response choices for prompts 1, 3 and 4 ranged from ‘not well / not easy at all’ to extremely well / extremely easy”, while response choices for prompt 2 ranged from ‘none at all’ to ‘all of it’. Potentially, similarity between prompt response options may encourage greater consistency in selected responses.

Table 3.3

Kendall’s Tau-b Between Judgement Prompt Responses Across Conditions.

Perceived Comprehension Prompt	1.	2.	3.
1. Overall, how well do you understand the text?	-		
2. How much of the text do you feel you understand?	0.64	-	
3. How easy was it to understand the text?	0.75	0.57	-
4. How well do you think someone else would understand the text?	0.71	0.60	0.76

Design Analysis. To evaluate the capacity to estimate the effect of interest with accuracy and precision in Study 1, an analysis of the proposed study design was conducted under varying sample sizes. As described in Chapter 2, achieving accuracy and precision in estimation was operationalised as obtaining an effect estimate which is simultaneously within $\pm\,\delta$ from the population-level effect and associated with a credible interval for the effect of width $\leq w$. The sample size of Study 1 is considered adequate when the probability of achieving the target level of accuracy and precision in estimation is expected to occur in at least 80% of hypothetical studies. Given the complexity of the study design and intended analysis, alongside the lack of an available closed-form solution, a simulation-based approach was taken to estimate the probability of achieving this goal of estimation.

The two research questions, RQ1 and RQ2, in Study 1 concern three parameters of interest: i) the effect of perceived comprehension, ii) the interaction between perceived comprehension and reading ability, iii) and the interaction between perceived comprehension and background knowledge. Given the considerable uncertainty, particularly the limited evidence for the direction of the interaction effects, it was considered that the Study 1 should primarily have the capacity to estimate the main effect of perceived comprehension with accuracy and precision. To this end, a multilevel logistic regression model was used to simulate response data to estimate the probability of obtaining an accurate and precise estimate for the slope coefficient for perceived comprehension. Selection of model parameter values were informed by Pilot 1 data and previous research, where appropriate. Given considerable uncertainty, greater variance magnitudes, with a conservative effect magnitude, were selected.

Assessed Comprehension Simulation Model. In the simulations, the binary response *Y* denotes whether or not an individual participant *i*, reading a health-related text j, answered a comprehension question *k* correctly ($Y_{ijk} = 1$) or incorrectly ($Y_{ijk} = 0$). Letting P_{ijk} be the corresponding probability of observing a correct response $P(Y_{ijk} = 1)$, this event can be expressed as the result of a Bernoulli trial:

$$Y_{ijk} \sim \text{Bernoulli}(P_{ijk}),$$

where *i* = 1, …, *I*, *j* = 1, …, *J* and *k* = 1, …, *K*; and *I*, *J* and *K* refer to the total number of participants, texts and questions per text, respectively. Assuming a logit-link function, the explanatory variables of interest and hierarchical sources of variance were linked to the response:

$$\log\left(\frac{P_{ijk}}{1 - P_{ijk}}\right) = \beta_0 + (\beta_1 + u_{1i})Judgement_{ij} + u_{0i} + v_j + w_k, \qquad (7)$$

where $Judgement_{ij}$ refers to the observed rating of perceived comprehension from individual *i* in response to text *j*; β_0 refers to the intercept (baseline log odds of observing success); β_1 refers to the population-level effect of perceived comprehension; u_{1i} refers to individual-level variability in the effect of perceived comprehension; and u_{0i}, v_j and w_k refer to intercept variability at the level of the individual, text and question, respectively.

Based on pilot data, pooled across text presence and response formats, the probability of answering a comprehension question correctly given the lowest rating of perceived comprehension (the baseline) was estimated to be around 0.1. A higher baseline probability was selected to capture the point at which there is maximum uncertainty in the response to the comprehension question: a baseline probability of 0.5 was selected. Solving the logit function yields $\beta_0 = 0$.

Additional sources of variance on the baseline probability of answering a comprehension question correctly were added to reflect the multilevel structure of the design Greater magnitudes of variability than suggested by the pooled pilot data were selected for these parameters. Participant-level u_{0i}, text-level v_j, and question-level w_k variability was specified as:

$$u_{0i} \sim N(0, 1)\,,$$

$$v_j \sim N(0, 0.5)\,,$$

$$w_k \sim N(0, 0.5)\,.$$

A conservative magnitude for the effect of perceived comprehension β_1 was selected, whilst balancing the capacity to resource a sample size sufficient to permit directional conclusions for small effects. Note that the selected magnitude of β_1 does not preclude the theoretical relevance of an observed difference below this value. As the observed outcome is binary, β_1 is expressed as a change in the log-odds of answering a comprehension question correctly, given a single unit increase in the perceived comprehension judgement (a higher rating of comprehension). A magnitude of 0.2 was selected for β_1, assuming a linear effect on assessed comprehension. Given that $\beta_0 = 0$, this value of β_1 corresponds approximately to an increase from 0.5 to 0.55 in the probability of correctly answering a question when the rating of perceived comprehension is increased from the first ($Judgement_{ij}$ = 0) to the second ($Judgement_{ij}$ = 1) response.

Considerable between-participant variability (u_{1i}) in the population-level effect of perceived comprehension was incorporated. Individual-level variance from β_1 was specified as:

$$u_{1i} \sim N(0, 0.2)\,,$$

yielding a range of values for u_{1i} spanning approximately -1 to +1. This allowed individuals to exhibit a strong negative, positive, or no association between perceived and assessed comprehension. The individual-level intercept (u_{0i}) and slope (u_{1i}) variances were specified to be uncorrelated, given that there is no clear evidence that individuals who are more likely to answer a comprehension question correctly should also be likely to show a stronger positive (or negative) association between perceived and assessed comprehension.

For $Judgement_{ij}$, a vector of observations for each individual *i*, of length *J*, was separately simulated as the sum of 4 Bernoulli trials per text, corresponding to the perceived comprehension judgements for each individual across the set of texts. To produce a similar pattern of ratings to those observed in the pilot data, the expected probability of each Bernoulli trial was 0.8. Variability from the probability of success on each trial was simulated per participant *i*, distributed $N(0, 1)$, and per text *j*, distributed $N(0, 0.5)$. This generated five possible values for $Judgement_{ij}$ = (0,1,2,3,4), corresponding to each of the five rating responses available for perceived comprehension judgements.

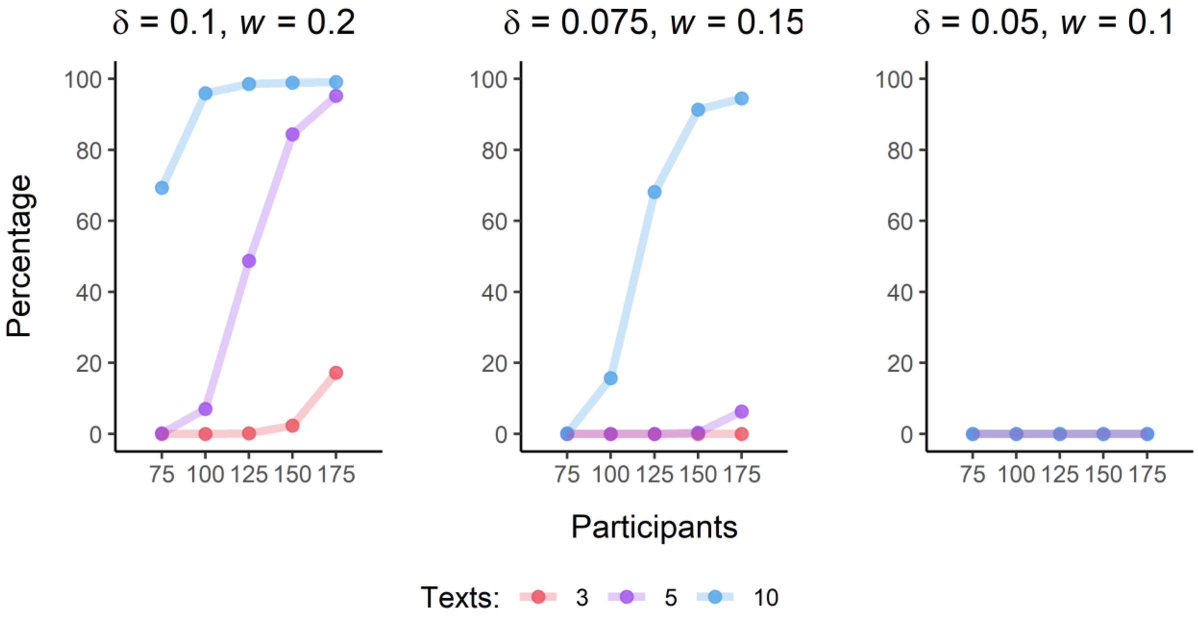
Simulation Procedure. The simulation was conducted using the High-End Computing (HEC) facility at Lancaster University using R (R Core Team, 2019). The simulation was conducted under varying numbers of participants and texts. Five values were considered for the total number of participants *I* = (75, 100, 125, 150, 175), and three values were considered for the total number of texts *J* = (3, 5, 10). Across all simulations, the total number of questions per text K was fixed to 5. A dataset of observations for each combination of participant and text sample sizes was simulated 1000 times.

Model Fitting. The multilevel logistic model defined in (7) was fitted to each simulated dataset to estimate the effect of perceived comprehension, using a Bayesian estimation framework, via the brms package (Bürkner, 2017, 2019) and Stan (Carpenter et al., 2017). Weakly informative priors specified for the model parameters. For the population-level effects (β_0, β_1), normal distributions with mean 0 and standard deviation 10 were specified. Half-student-t distributions with the values of the degrees of freedom, location and scale parameters set to 3, 0 and 10 for the group-level effects (u_{0i}, u_{1i}, v_j, w_k). An LJK correlation distribution with a shape parameter of 1 was used as the prior on the covariance between participant-level variance in the intercept and the effect of perceived comprehension. This prior allows any correlation matrix to be equally probable a priori (Lewandowski, et al., 2009). From each resulting model fit, estimates for the mean of the posterior distribution and the 90% credible interval for the effect of interest were obtained.

Simulation Results. The capacity to estimate the effects of interest, for a given number of participants and texts, was operationalised in terms of the probability of estimating β_1 with accuracy and precision. Three pairs of values were selected for δ and w representing **lower** ($\delta = 0.1, w = 0.2$), **middle** ($\delta = 0.075, w = 0.15$) and **higher** ($\delta = 0.05, w = 0.1$) levels of accuracy and precision in estimation. To better understand how these values for δ and w relate to observed effect estimates, the potential values which may be observed for effect estimates and the values contained within the credible intervals can be considered. For the **lower** level of accuracy and precision, successful estimation of the effect excludes point estimates of less than half of the true effect magnitude and, separately, excludes values of more than double the true effect magnitude from the 90% credible intervals. For the **higher** level of accuracy and precision, successful estimation of the effect simultaneously excludes values of less than half and more than double the true effect magnitude from the 90% credible intervals. The **middle** level of accuracy and precision represents a mid-point between these two estimation outcomes. Across each of selected sets of values for δ and w, successful estimation of the effect produces a 90% credible interval which does not overlap with zero. The probability of successfully achieving accuracy and precision in estimation, across varying δ and w, for each total number of participants and texts simulated, is shown in Figure 3.1.

Figure 3.1

Probability of Achieving Varying Levels of Accuracy and Precision in Estimation in Pilot 1 Simulation



Note. The lines show the percentage of simulations achieving the specified target level of accuracy (δ) and precision (w) under varying numbers of participants (horizontal axis) and numbers of texts (coloured lines).

From left to right in Figure 3.1, decreasing values of δ and w correspond to higher demands on accuracy and precision. The simulation indicated that, across the values of δ and w considered, increasing the number of participants while the total number of texts remained small did not have a considerable benefit on the probability of estimating with accuracy and precision. Given the **lower** level of accuracy and precision selected (left plot in Figure 3.1), an 80% probability of achieving accuracy and precision occurs with at least 100 participants responding to 10 health-related texts. In contrast, the probability of achieving the **higher** level of accuracy and precision selected, across the numbers of participants and texts simulated, is close to zero (right plot in Figure 3.1). For the **middle** level of accuracy and precision (middle plot in Figure 3.1), adequate probability of achieving accuracy and precision only occurs when the total number of texts is 10.

3.2.2 Pilot 2

The simulation-based analysis of the design of Study 1 suggested that the proposed design (using three health-related texts) was unlikely to yield estimates for the effect of interest which met the examined levels of accuracy and precision in estimation, even with many participants. Increasing the number of texts was found to offer a considerable gain in the accuracy and precision of effect estimation. Given the length of stimulus texts in the proposed design, increasing the number of texts would considerably extend the duration of the experiment, potentially leading to participant fatigue.

To avoid participant fatigue, a number of factors were considered to reduce testing duration while accommodating greater numbers of texts. This included shortening the length of the stimulus texts, presenting comprehension questions in multiple-choice response format, and lowering the number of questions per text. Based on approximate task durations observed in the first pilot study, a similar length of testing session could be achieved using 10 texts of 300 words each, with four questions per text. A second pilot (Pilot 2) was undertaken to test this alternative design and inform an assessment of the required sample size to estimate the effect of interest with accuracy and precision.

In addition to examining the alternative proposed number of shorter texts, the second pilot was used to test the implementation of the selected task formats to be used in Study 1. To measure reading ability, the life cycle of stars text was selected and administered in open-ended response format with the text present. This format was selected due to the comparable performance on open-ended responses between the two texts examined and to maintain consistency with previous applications of the QRI. For the measure of background knowledge, the HLVA was administered in multiple-choice question format. This format was selected given potential issues with participants providing definitions informed by online searching and the benefit of reduced time required to complete the task.

The comprehension questions for the metacomprehension task were presented in multiple-choice response format to reduce the time required for participants to complete these. This was also considered reasonable as performance did not appear to differ substantially between response formats in Pilot 1. The questions were presented with the health-related text present. This was to reduce the additional demands on memory introduced by requiring participants to read and judge nine other texts before answering comprehension questions. While the use of multiple prompts for metacomprehension was anticipated in Study 1, only two prompts were used in the second pilot to consider how consistency in prompt response options may influence the associations between metacomprehension judgements.

Participants. A sample of 10 UK nationals was recruited using the online platform Prolific. Participants were aged 19 to 64 (*M* = 37.1, *SD* = 15.04), with 8 identifying as female and 2 as male. Participation was restricted to those whose highest level of qualification did not exceed A-Level standard. Participants received £8.00 compensation for taking part.

Materials.

Qualitative Reading Inventory (QRI). Reading ability was measured using the text concerning the life cycles of stars. Responses to the 20 QRI questions, administered in the open-ended response format, were summed to creating a reading ability score.

Health Literacy Vocabulary Assessment (HLVA). Topic-specific vocabulary knowledge was measured using the multiple-choice response option version of the HLVA. The same test items and response options from the first pilot were used. Responses were summed to create a background knowledge score.

Health Texts. A set of ten stimulus texts were created using information from NHS Choices online A-Z of health conditions, obtained via opportunity sampling. Ten topics were selected which were of similar composition and content, were considered unlikely to be highly familiar to participants, and which did not concern potentially embarrassing conditions. Approximately the first 300 words of each health condition were used as a stimulus text. As previously, headers and font formatting were preserved, but sign-posting text, links and images were removed. The resulting 10 stimulus texts are provided in Appendix E, and descriptive information is provided in Table 3.4.

Consistent with the difficulties discussed by others (Griffin, Mielicki & Wiley, 2019), identifying adequate material to develop situation-model level inference-based questions within shorter texts was challenging. Previously, suitable questions were identified by searching the stimulus texts to identify two pieces of information that could be connected via an inference which was amenable to question construction. Due to the limited volume of information within the texts, an alternative approach was taken which first involved considering what constituted successful comprehension of each text. To identify the main points of information a reader would be expected to successfully comprehend after reading, stimulus texts were evaluated in two ways. Firstly, for each text, a gist-level summary was written and the main points of information in the text were listed. In addition, a second researcher, naive to the purpose of the task, read each text and provided an oral summary. From these summaries, important elements within each text were identified for question construction.

Table 3.4

Descriptive Information for Health Texts in Pilot 2

Topic	Words	Flesch Score	Flesch-Kincaid
Aspergillosis	316	54.6	9.8
Autosomal dominant polycystic kidney disease	282	49.9	11.6
Non-melanoma skin cancer	315	53.8	10.1
Behcet's disease	304	37.4	14.3
Schistosomiasis	289	59.0	10.1
Bornholm disease	302	55.4	9.5
Brugada syndrome	325	55.0	9.6
Transient ischaemic attack	302	68.2	8.7
Cholesteatoma	291	58.9	10.2
Isovaleric acidaemia	316	47.7	10.4

Note. Flesch score refers to the Flesch Reading Ease score. Flesch-Kincaid refers to the Flesch-Kincaid Grade level. Calculated using the online tool Coh-Metrix 3.0 (Graesser, et al., 2004).

Key points of information identified from the stimulus texts varied in the extent to which they were amenable to inference-based questions. For the majority of questions, correct responses questions captured generative inferences made on the basis of successful comprehension of the information. However, this was not possible for all questions, particularly those concerning explicit details (e.g., that the cause of a disease was a virus).

For these questions, correct responses captured the meaning of the stated information, whilst minimising lexical overlap with the text as much as possible.

Given variability in the construction of questions, it was challenging to consistently construct response options for each question according to a defined set of distractor categories. Consequently, distractors were less systematically created in Pilot 2. Distractors primarily consisted of words or phrases presenting semantic near-misses, high lexical overlap with the text, or likely misconceptions based on prior knowledge (near-miss inferences). The four comprehension questions for each text, with the multiple-choice responses, are provided in Appendix F.

Perceived Comprehension. The two prompts which appeared to produce the least similar metacomprehension judgements in Pilot 1 were used in Pilot 2: ‘Overall, how well do you understand the text?’ and ‘How much of the text do you feel you understand?’. The rating response scales were consistent with Pilot 1.

Procedure. Participation in the task commenced when participants responded to the online invitation on Prolific.co. All tasks were computer-based and presented via Qualtrics. Participants provided consent and reported their age and gender prior to completing the tasks. The instructions and order of task presentation was the same as in Pilot 1.

Results and Discussion. Pilot 2 was undertaken to evaluate the adapted design for Study 1. Given the small sample size, variation between conditions should be treated as weakly indicative of differences. Performance on the QRI life cycle of stars text was consistent with that observed in Pilot 1 (*M* = 12.4, *SD* = 5.23). Performance on the HLVA was also similar to that observed in Pilot 1 (*M* = 7.5, *SD* = 1.65). In contrast, average performance on the multiple-choice questions relating to the health-related texts in Pilot 2 (*M* = 2.94, *SD* = 1.02) was greater than Pilot 1. As an average proportion of the total sum score per text, this reflects an increase from 36% to 74%. This may be due to the different methods of deriving comprehension questions, the response options employed, or individual variability.

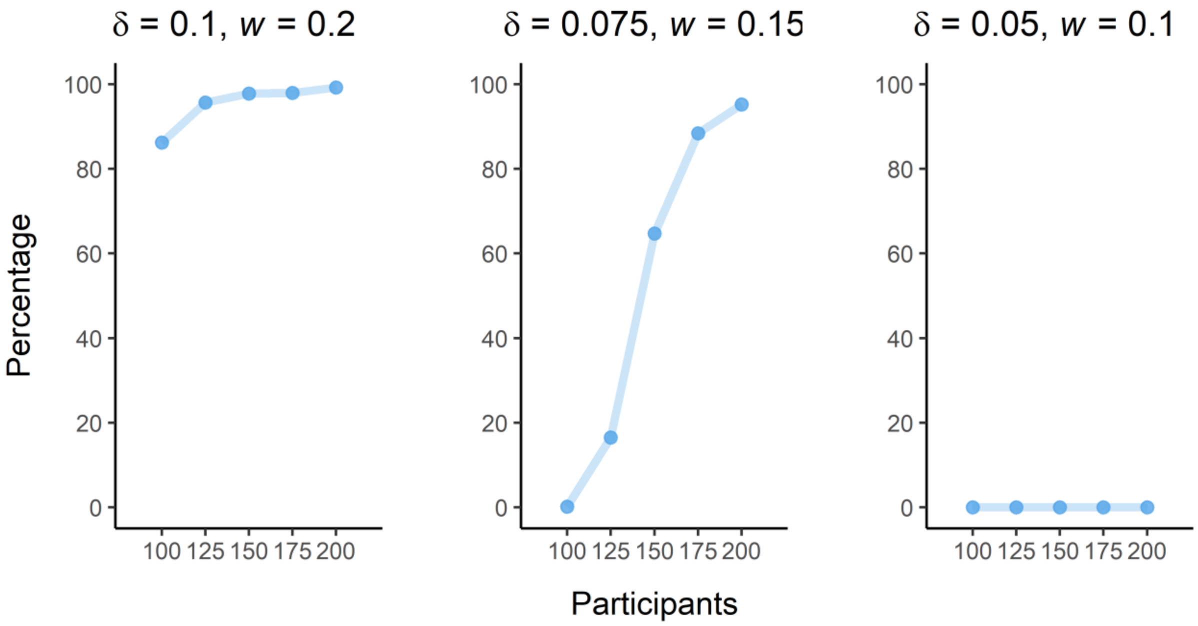
The correlation between the perceived comprehension prompts demonstrated stronger alignment than in Pilot 1. The correlation between responses to ‘Overall, how well do you understand the text?’ and ‘How much of the text do you feel you understand?’ was greater: Tau-b = 0.84. This suggests that metacomprehension judgements may be influenced by the presence of additional prompts.

Design Analysis. To determine the sample size required to estimate with accuracy and precision in Study 1, given the revised design, a second simulation conducted. The model and parameterisation for the simulations was largely similar to that described in the design analysis following Pilot 1 (section 3.2.1), but with the following changes. The baseline probability of answering a comprehension question correctly was increased to 0.55 to reflect the higher performance on the comprehension questions observed in Pilot 2. Solving the logit function yields β_0= 0.2. Due to the change in β_0, the effect of perceived comprehension β_1 was adjusted to 0.205. This was to maintain consistency with the selected smallest effect of interest, corresponding to an increase of 0.05 in probability of correctly answering a comprehension question, given an increase in the perceived comprehension judgement from the first ($Judgement_{ij}$ = 0) to the second ($Judgement_{ij}$ = 1) response. The total number of questions per text K was reduced to four and the total number of texts J was fixed at 10. Five values were simulated for the total number of participants I = (100, 125, 150, 175, 200). All other aspects of the simulation, including model fitting, remained the same.

The same three pairs of values for the target levels of accuracy and precision were selected, representing **lower** ($\delta = 0.1, w = 0.2$), **middle** ($\delta = 0.075, w = 0.15$) and **higher** ($\delta = 0.05, w = 0.1$) levels of accuracy and precision in estimation. The probability of successfully achieving these levels of accuracy and precision in estimation, for each total number of participants simulated, shown in Figure 3.2, was highly consistent with that observed in the previous simulation. Given **lower** levels of accuracy and precision, an adequate probability of achieving the goal of estimation was observed in all sample sizes (left plot, Figure 3.2), whereas none of the sample sizes considered satisfied the **higher** target level of accuracy and precision (right plot, Figure 3.2). A larger sample size was required to yield an adequate probability of achieving the **middle** level of accuracy and precision compared to the first simulation (middle plot, Figure 3.2). Given five questions per texts, a sample of 150 participants was estimated to yield a 91% probability of achieving the **middle** level of accuracy and precision in the first design simulation. Reducing the number questions per text to four, however, indicated that 175 participants would yield an estimated 88% probability of achieving this goal. This difference reflects the lower number of observations per participant, hence the lower level of information furnished in this design.

Figure 3.2

Probability of Achieving Varying Levels of Accuracy and Precision in Estimation in Pilot 2 Simulation



Note. The lines show the percentage of simulations achieving the specified target level of accuracy (δ) and precision (w) under varying numbers of participants (horizontal axis).

Based on the simulations under the revised study design, ten texts with four questions per text would provide a design sufficient to estimate the effect of interest with accuracy and precision, conditional on the number of participants. Balancing a number of concerns, including the desired level of accuracy and precision in estimation, available resources and uncertainty in the simulation, a sample of 175 participants was chosen for Study 1. This sample would yield approximately an 88% probability of achieving the **middle** level of accuracy and precision in estimation.

3.2.3 Participants

In Study 1, a sample of 175 participants was recruited using the online platform Prolific (prolific.co). Participation was limited to UK nationals that had not taken part in the pilot studies. Recruitment was conducted over several days to capture a greater spread of participants on the Prolific system. Exclusion criteria, agreed prior to data collection, specified that participants would be excluded and replaced if they demonstrated limited evidence of engagement with the task. Reading times were used as a proxy measure of engagement. Exclusion based on reading times was designed to ensure that the measures of perceived and assessed comprehension were informed by understanding gained from the reading text.

To select the reading time exclusion criteria, it was considered that participants should be minimally expected to read the whole text as instructed. The required time for an individual to read the entire text could, therefore, be estimated based on the length of the text. However, individual reading rates vary between individuals, reflecting slower or faster readers, and within individuals, reflecting the type of reading engaged in. Carver (1992) proposed a typology of reading 'gears' which reflect different reading goals and are associated with different reading rates. Carver (1992) suggested that the reading rates for normal silent reading, skim-reading, and scanning, measured in words per minute (wpm), are 300, 450 and

600, respectively. More recently, a meta-analysis of 190 studies indicated that the average English silent reading rate for adults reading non-fiction is 238 wpm, with the reading rates of most adults falling within the range of 175-300 wpm (Brysbaert, 2019). Limited evidence for Carver's (1992) reading categories was found, apart from the distinction between normal reading and scanning. Further, Brysbaert (2019) suggested that there may be a difference between skimming and scanning, in which skimming is a dynamic mix of normal and scanning reading behaviours, but that further research was required to robustly evidence this distinction.

In Study 1, it was considered that individuals may engage in some form of skimming in their natural approach to consuming novel health-related information, therefore this style of reading was permissible. As the distinction between the reading rates of skimming and scanning are not clear, the higher rate of 600 wpm was selected (Carver, 1992). This reading rate was selected to permit considerable individual variability in skim reading behaviours whilst excluding individuals who show limited evidence of engaging with the text beyond scanning reading behaviours. Based on this reading rate and the word length of stimulus texts (*M* = 300.8, *SD* = 13.77), participants were excluded if the time spent on any of the texts was less than 30 seconds. An evaluation of the impact of exclusion on the results are provided in the sensitivity analysis (section 3.3.3).

Following the sampling procedure outlined, participants were recruited until 175 submissions which met the acceptance criteria were obtained (page submissions times of ≥ 30 seconds on all 10 health texts). In total, 255 participants were recruited. Responses from 70 participants who met the exclusion criteria and 10 participants who did not fully complete the study were excluded and replaced. The sample of accepted participants (*I* = 175) consisted of 109 female and 66 male participants, with age ranging from 18 to 76 (*M* = 37.78, *SD* = 12.91). All participants who fully completed the study received a payment of £5.00.

3.2.4 Materials

The materials used to capture the four variables of interest (perceived comprehension of a text, assessed comprehension of a text, reading ability and background knowledge) were as described in the second pilot study, with some minor differences.

QRI. Comprehension performance on the two ‘life cycle of stars’ passages, taken from the QRI, provided a measure of reading ability. Participants provided open-ended responses to the 20 questions. Marking of responses deviated from the pilot studies: an altered version of the QRI marking rubric was developed following full data collection in Study 1. In the context of the large variability observed in participant responses to the questions, the marking criteria were considered somewhat unclear and therefore difficult to implement whilst reliably identifying answers which were correct. To address this uncertainty, the marking rubric was supplemented with additional examples to clarify what constituted acceptable and insufficient responses. Two experimenters marked responses using the altered rubric. Marker agreement was generally high (percentage agreement per question ranged from 81% to 100%, *M* = 96%). Each disagreement was discussed and resolved jointly.

HLVA. The 16-item adaptation of the HLVA, with full-scoring applied, provided a measure of health-related background knowledge. Responses were collected via multiple-choice questions, identical to those used in Pilot 2.

Health Texts. The 10 health texts used in the metacomprehension task were identical to those used in Pilot 2. Minor changes to a small number of the question prompts and answer options used were made on the basis of reviewing the Pilot 2 data. The altered questions are provided in Appendix G.

Perceived Comprehension. To allow for comparisons of the estimated relationship between metacomprehension judgements and assessed comprehension between judgement prompts, five prompts were used to elicit judgements. Three of these were selected from Pilot

1 and 2: ‘Overall, how well do you understand the text?’, 'How much of the text do you feel you understand?', and ‘How easy was it to understand the text?’. Responses to two additional prompts were collected, based on common questions reader panel members are asked during health-text evaluations: ‘How well was the text written? (In terms of spelling, punctuation and grammar)’, and ‘How patient-friendly was the text? (In terms of technical or medical words)’.

From the judgement prompts, one was selected for use as the measure of perceived comprehension in the main analyses, while the remaining four were used in a sensitivity analysis to examine the robustness of the main analysis to differences in the prompts. The second prompt 'How much of the text do you feel you understand?' was selected, given the apparent similarity between prompts, potentially conditional on the presence of other prompts, observed in the pilot studies. Judgements were captured using rating scales described in Pilot 1 and 2. For the two additional prompts, response options were presented as 'not well at all/not at all’, 'slightly well/slightly', 'moderately well/moderately', 'very well/very' and 'extremely well/extremely' for the prompts concerning the writing-style and language accessibility, respectively.

3.2.5 Procedure

Participation in the task commenced when participants responded to the online invitation on Prolific.co. All tasks were computer-based and presented via Qualtrics. Participants provided consent and reported their age and gender prior to completing the tasks. The instructions and order of task presentation was the same as Pilot 1. Participants first completed the HLVA, followed by the metacomprehension task. Participants read each health-related text and made judgements of comprehension immediately following the reading of each text. Participants were invited to take a break if needed after reading and judging all texts. Participants were then presented with the comprehension questions for each text, with the text present, in the same order that the texts were presented for reading. Lastly, participants read and completed the questions, with the text present, for the two passages of the life cycle of stars text from the QRI.

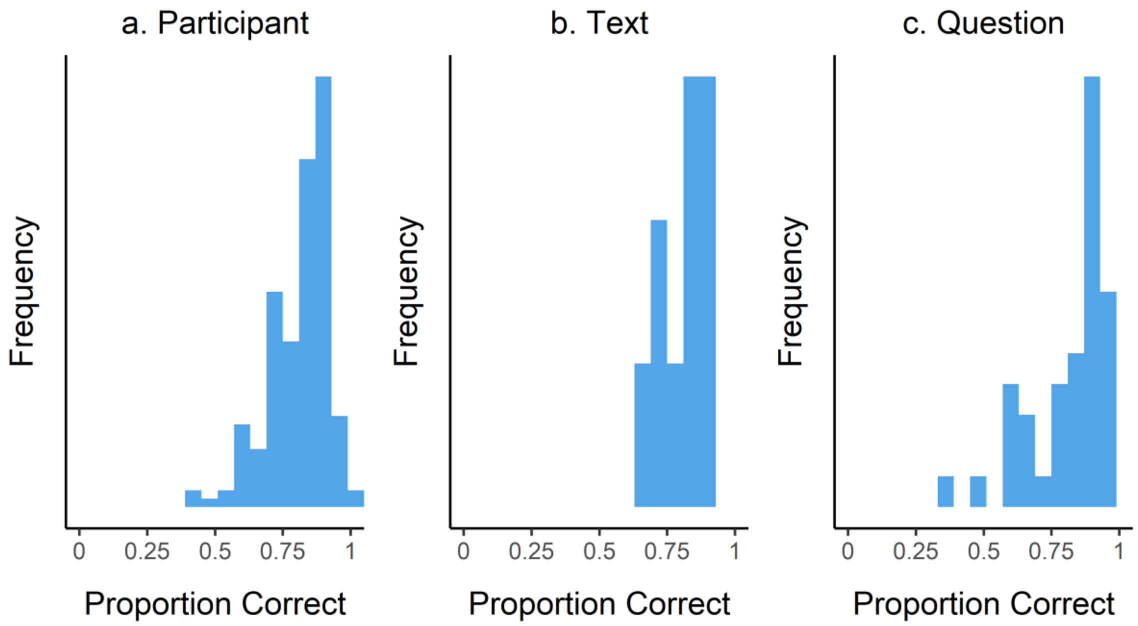
3.3 Results

3.3.1 Preliminary Data Inspection

Initial inspection of the responses which met the acceptance criteria indicated that performance on comprehension questions was generally high. The proportion of comprehension questions answered correctly, across all participants, texts and questions, was .78. This corresponds to participants answering approximately three of the four questions per text correctly, on average. Variability in this was observed at the level of the participant, text and question. Figure 3.3 shows the distributions of the proportion of correct responses by participant (3.3a.), text (3.3b.), and question (3.3c.). Variability in proportion correct was comparably greater between both participants and questions than texts.

Figure 3.3

Proportion of Comprehension Questions Correctly Answered in the Metacomprehension Task in Study 1



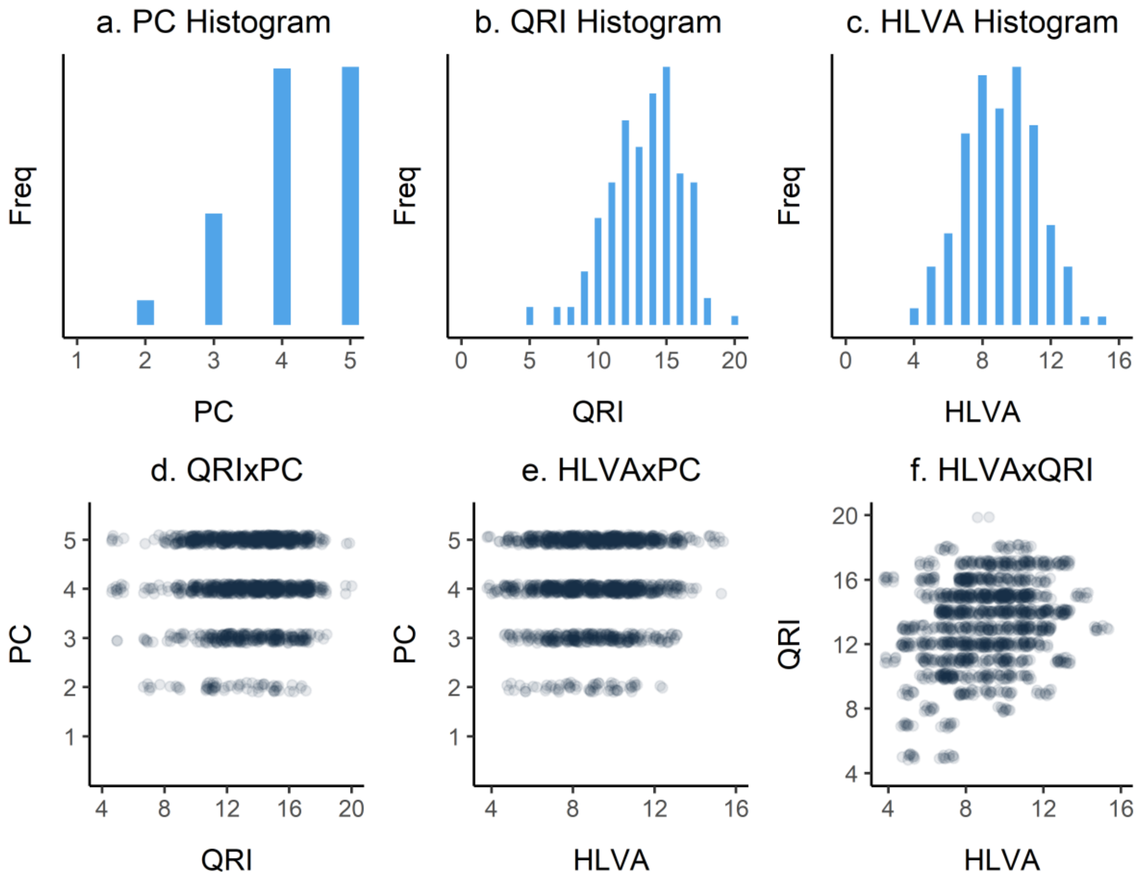
Note. Participant proportion correct (*I* = 175) is calculated as the total questions correctly answered by the participant divided by the total number of questions. Text proportion correct (*J* = 10) is calculated as the total questions correctly answered per text divided by the product of the number of questions per text and number of participants. Question proportion correct (*K* = 40) is calculated as the number of participants correctly answering the question divided by the total number of participants.

The distributions of variables to be used as predictors in the main analysis (perceived comprehension judgements, QRI scores and HLVA scores) and their corresponding bivariate scatterplots are shown in Figure 3.4. Responses to the perceived comprehension prompt selected for use in the main analysis: ‘How much of the text do you feel you understand?’ were negatively skewed. As can be seen in Figure 3.4a, participants predominantly selected the response options 'a lot' and 'all of it'. Differences in judgement distributions were observed between texts (see Figure H.1 in Appendix H). Contrary to expectations based on piloting, no respondents selected the response option ‘none of it’ on any text. The scatterplots of perceived comprehension and both the QRI scores (3.4b.) and HLVA scores (3.4c.) indicated no clear bivariate association between these measures.

Across participants, performance on both the QRI (*M* = 13.39, *SD* = 2.61) and the HLVA (*M* = 9.04, *SD* = 2.13), Figures 3.4b and 3.4c, respectively, were broadly normally distributed, though performance on the QRI showed slight negative skew. These distributions indicate that the sample captured a range of different levels of reading ability and health-related background knowledge. The scatterplot between these variables, shown in Figure 4f, suggested a weakly positive association.

Figure 3.4

Histograms and Scatterplots of the predictor variables considered in Study 1



Note. Points in scatterplots d.-f. are plotted with jitter applied, adding small, randomised perturbations to the observations to increase the readability of the plots.

3.3.2 Planned Analysis

To explore whether judgements of comprehension are predictive of assessed comprehension of health-related texts and whether individual differences in reading ability and background knowledge might moderate such a predictive relationship, Bayesian multilevel logistic regression models were fitted in R (R Core Team, 2019) using the brms package (Bürkner, 2017, 2019) and Stan (Carpenter et al., 2017). The results of the analyses corresponding to each research question are discussed below. Following this, a sensitivity analysis is reported to explore how various analytical choices may influence the findings.

RQ1. To address RQ1, the model defined in (7), repeated below for convenience, was fitted to the data using a Bayesian estimation framework:

$$\log\left(\frac{P_{ijk}}{1-P_{ijk}}\right) = \beta_0 + (\beta_1 + u_{1i})Judgement_{ij} + u_{0i} + v_j + w_k\ . \qquad (7)$$

Observations of assessed comprehension, from individual *i* on text *j* and question *k*, were used as the outcome variable Y_{ijk}. Judgements of perceived comprehension, standardised prior to model fitting, were entered as predictors of the log odds of correctly responding to the comprehension question. Group-level variance in the intercept was estimated for participants u_{0i}, texts v_j and questions w_k. Participant-level variability in the effect of perceived comprehension u_{1i} was included. The same weakly informative priors described in the design analysis (section 3.2.1) were used for the population-level effects, group-level effects, and covariance parameters.

The model was estimated using six chains with 8000 iterations each, half of which were discarded as burn-in. Due to difficulties in exploring values close to zero in participant-level variability in the effect of perceived comprehension, the ‘adapt_delta’ parameter in Stan was set to 0.99 to resolve the small proportion of divergent transitions. This parameter reduced the step size of the sampler, leading to a fuller exploration of the sampling space and the production of a more robust posterior distribution. Model convergence was evaluated by inspecting indices of convergence, including estimates of potential scale reduction factor, autocorrelation and effective sample size, alongside visual assessment of posterior distributions and trace plots of sample chains. No issues were indicated, suggesting that the model appeared to converge well under this specification. The results of this model are presented in Table 3.5.

Table 3.5

Bayesian Multilevel Logistic Model of Response Accuracy in Study 1

Parameter	Estimate[a]	Error[b]	95% CI[c]	Eff Sample
Population-Level Effects				
Intercept	1.90	0.24	[1.43, 2.37]	7556
Perceived comprehension	0.12	0.05	[0.03, 0.21]	25108
Group-Level Variance				
Participant (intercept)	0.73	0.06	[0.62, 0.85]	9063
Participant (perceived comprehension)	0.07	0.06	[0.00, 0.21]	5454
Text (intercept)	0.40	0.28	[0.02, 1.04]	2738
Question (intercept)	1.07	0.14	[0.82, 1.39]	6742
Covariance of intercept and slope variance				
Participant intercept-slope correlation	0.12	0.49	[-0.88, 0.93]	22656

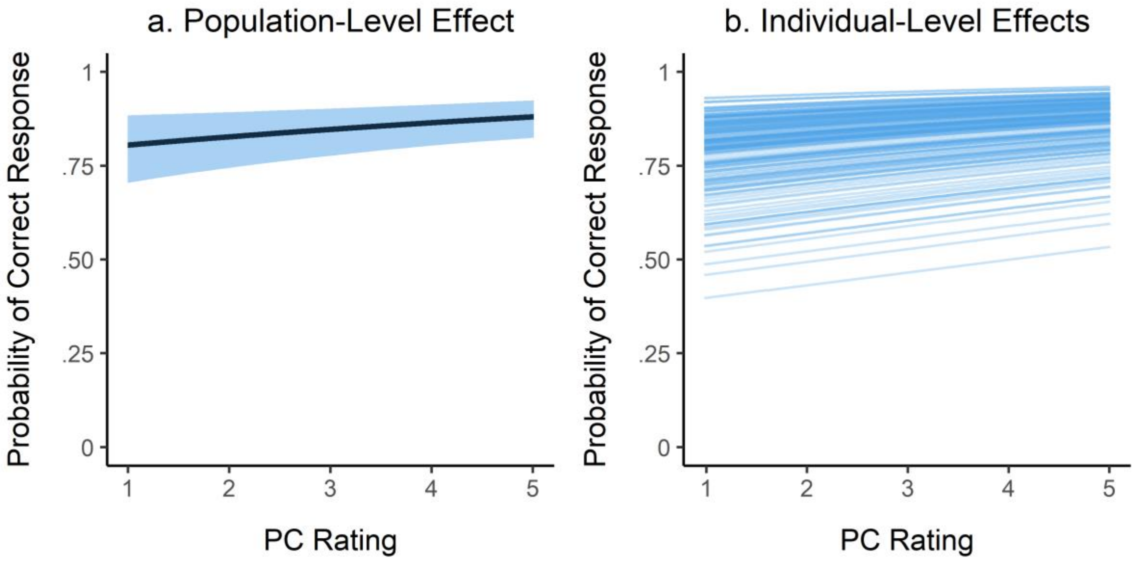
Note: Population-level effect estimates are presented in logits. Rhat values for all parameters = 1.00. CI = credible interval. Eff Sample = number of effective samples.

[a] Estimate refers to the mean of the marginal posterior distribution of the parameter. [b] Error refers to the standard deviation of the marginal posterior distribution of the parameter. [c] Credible intervals represent the upper and lower values within which 95% of the estimated parameter values in the posterior distribution are contained.

The mean of the posterior distribution for perceived comprehension judgements was estimated to be positive, indicating that the log odds of answering a comprehension question correctly increases on average by 0.12 per unit increase in perceived comprehension. A considerable portion of the posterior probability density was located across positive values: both the 95% credible interval (see Table 3.5) and the 80% highest density interval (80% HDI = [0.06, 0.18]) provide certainty in the direction of the effect. To illustrate the estimated population-level effect of perceived comprehension, predictions of the probability of answering a comprehension question correctly for a given rating of perceived comprehension can be calculated. Based on model-fitted predictions of the marginal effect, a perceived comprehension judgement of understanding 'none at all' compared to understanding 'all of it' would correspond to a probability of answering a comprehension question correctly of 0.81 (*SE* = 0.05) and 0.88 (*SE* = 0.03), respectively. This estimated increase is illustrated in Figure 3.5a, with the 95% credible interval shaded to illustrate the uncertainty in this effect.

Figure 3.5

Estimated Effect of Perceived Comprehension, at the Population- and Individual-Level, in Study 1



Note. Marginal model-fitted predictions at the population-level (a.) and conditional model-fitted predictions at the individual-level (b.) are plotted. Panel a. shows the expected probability of a correct response for the 'average' participant, responding to the 'average' question concerning an 'average' text, with the 95% credible interval shaded. Panel b. shows the expected probability of a correct response for each participant, given an 'average' question concerning an 'average' text. PC = perceived comprehension.

Participant-related variability in the predictive relationship between perceived and assessed was found to be limited. Indeed, the sampler had difficulties exploring this parameter space due to the close-to-zero variance (lower bound of the 95% credible interval = 0.003). This indicated that the individuals do not markedly differ from the population-level effect of perceived comprehension. Model-fitted estimates of the effect of perceived comprehension, for each participant, are shown in Figure 3.5b. The near parallel lines in this figure demonstrate that the relationship remains comparably small and positive across individuals. In addition, the estimated covariance between participant-related variability in answering a comprehension question correctly and participant-related variability in the effect of perceived comprehension was small with considerable uncertainty (see Table 3.5).

Differing magnitudes of group-level variability in the overall probability of answering a comprehension question correctly were estimated across participants, texts and questions. The largest source of variability was estimated at the level of questions, followed by participants, then texts. This reflects the observed variability response accuracy (Figure 3.3), indicating that heterogeneity in question and individuals were the main sources of group-level variability in response accuracy. Lower variability in response accuracy was found between.

RQ2. To address RQ2, the regression model defined in (7) was extended to estimate the effects of QRI score β_2, HLVA score β_3, and the interactions between QRI score and perceived comprehension β_4 and HLVA score and perceived comprehension β_5:

$$\begin{aligned}\log\left(\frac{P_{ijk}}{1-P_{ijk}}\right) &= \beta_0 + (\beta_1 + u_{1i})Judgement_{ij} + \beta_2 QRI_i + \beta_3 HLVA_i \\ &\quad + \beta_4 Judgement_{ij} QRI_i + \beta_5 Judgement_{ij} HLVA_i \\ &\quad + u_{0i} + v_j + w_k\ .\end{aligned} \tag{8}$$

Judgements of perceived comprehension, QRI scores and HLVA scores were standardised prior to model fitting. The same weakly informative priors described in the design analysis (section 3.2.1) were used for the population-level effects, group-level effects, and covariance parameters. For the four additional population-level effects ($\beta_2, \beta_3, \beta_4, \beta_5$), normal distributions with mean 0 and standard deviation 10 were specified as priors.

The model was estimated using six chains with 8000 iterations each, half of which were discarded as burn-in. Similar to the model previously fitted, the 'adapt_delta' parameter in Stan was set to 0.99 to resolve the small proportion of divergent transitions arising from difficulties in exploring values close-to-zero in participant-level variability in the effect of perceived comprehension. The marginal posterior distributions of the parameters did not markedly differ due to this change. Model convergence was evaluated as previously described. No issues were indicated, suggesting that the model appeared to converge well under this specification. The results of this model are presented in Table 3.6.

The estimated population-level effect of perceived comprehension on the probability of answering a comprehension question correctly was comparable to that obtained from model (7). The posterior mean for this effect indicated an expected increase of 0.13 in the log odds of correctly answering a comprehension per unit increase in perceived comprehension. Both the 95% credible interval (see Table 3.6) and 80% HDI for this effect contained only positive values (80% HDI = [0.07, 0.19]).

The main effects of QRI and HLVA were estimated to be positive, with the 95% credible intervals of these posterior distributions excluding zero and negative values. To illustrate the impact of these estimated effects, predictions of the probability of answering a comprehension question correctly can be calculated while holding the values of other predictors at their mean. Given the lowest and highest QRI scores observed in the data (5 and 20), the model-fitted predictions of the probability of answering a comprehension question correctly are 0.73 (*SE* = 0.05) and 0.93 (*SE* = 0.02), respectively. Given the lowest and highest HLVA scores observed in the data (4 and 15), the model-fitted predictions of the probability of answering a comprehension question correctly are 0.78 (*SE* = 0.05) and 0.93 (*SE* = 0.02), respectively.

Table 3.6

Extended Bayesian Multilevel Logistic Model of Comprehension Question Responses

Parameter	Estimate[a]	Error[b]	95% CI[c]	Eff Sample
Population-Level Effects				
Intercept	1.90	0.24	[1.43, 2.37]	7596
Perceived comprehension	0.13	0.05	[0.04, 0.22]	25126
QRI	0.28	0.06	[0.16, 0.39]	13407
HLVA	0.26	0.06	[0.15, 0.38]	12368
Perceived comprehension x QRI	0.07	0.04	[-0.01, 0.15]	25151
Perceived comprehension x HLVA	0.01	0.05	[-0.08, 0.10]	26763
Group-Level Variance				
Participant (intercept)	0.58	0.06	[0.48, 0.69]	10263
Participant (perceived comprehension)	0.08	0.06	[0.00, 0.21]	6028
Text (intercept)	0.40	0.28	[0.02, 1.05]	2566
Question (intercept)	1.07	0.15	[0.83, 1.39]	6127
Covariance of intercept and slope variance				
Participant intercept-slope correlation	-0.25	0.47	[-0.96, 0.79]	19604

Note: Population-level effect estimates are presented in logits. Rhat values for all parameters = 1.00. CI = credible interval. Eff Sample = number of effective samples. QRI = Qualitative Reading Inventory. HLVA = Health Literacy Vocabulary Assessment.

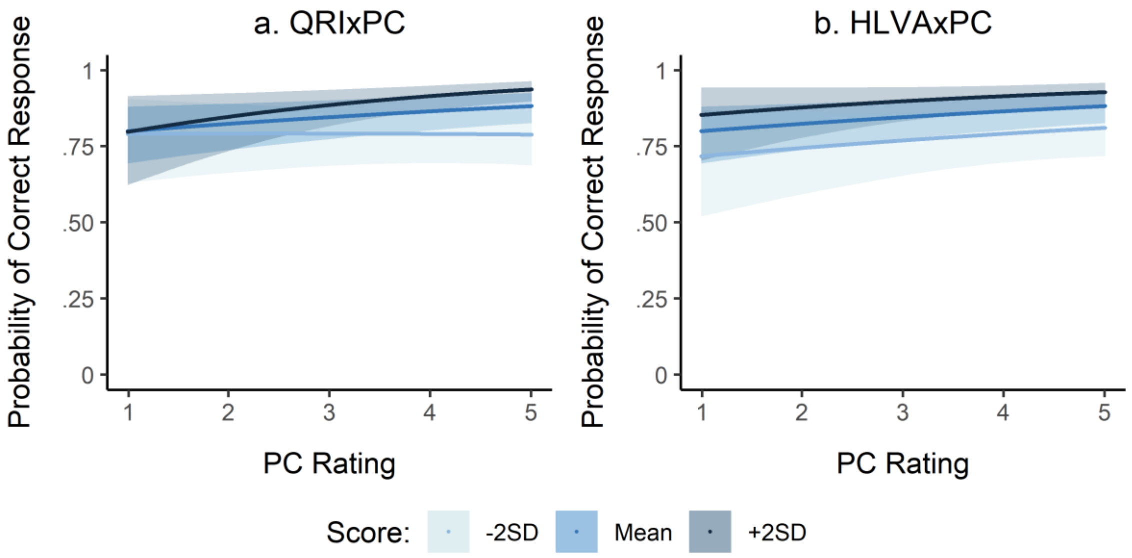
[a]Estimate refers to the mean of the marginal posterior distribution of the parameter. [b]Error refers to the standard deviation of the marginal posterior distribution of the parameter. [c]Credible intervals represent the upper and lower values within which 95% of the estimated parameter values in the posterior distribution are contained.

Limited evidence in support of an interaction between perceived comprehension and QRI score was found: the posterior mean for this effect was positive, however, the 95% credible interval included zero and negative values (see Table 3.6). An interaction in the direction indicated by the posterior mean would suggest that perceived comprehension judgements from individuals who score more highly on the QRI translates to greater increases in the probability of answering a question correctly than do judgements from individuals who score lower on the QRI. This is illustrated in Figure 3.6a, showing the model-fitted predictions of the effect of perceived comprehension across varying levels of performance on the QRI.

In contrast, no evidence for an interaction between perceived comprehension and HLVA score was found. The posterior mean for this effect was positive, however, the 95% credible interval was centred close to zero, including both positive and negative values (see Table 3.6). As illustrated in Figure 3.6b, across varying levels of performance on the HLVA, the model-fitted predictions of the effect of perceived comprehension remain equivalent.

Figure 3.6

Estimated Effect of Perceived Comprehension, at the Population-Level, by Performance on the QRI and HLVA, in Study 1



Note. Marginal model-fitted predictions at the population-level, across varying levels of performance on the QRI (a.) and HLVA (b.) are plotted. The panels show the expected probability of a correct response for the 'average' participant, responding to the 'average' question concerning an 'average' text, conditioned on QRI and HLVA performance, with the 95% credible interval shaded. Scores of -2SD and +2SD refer to scores two standard deviations below and above the mean score, respectively. PC = perceived comprehension, QRI = Qualitative Reading Inventory, HLVA = Health Literacy Vocabulary Assessment.

Estimates of group-level variability in the overall probability of answering a comprehension question correctly across participants, texts and questions remained largely consistent with the previous model (7). The inclusion of QRI and HLVA scores produced a slight reduction in participant-level variability in the probability of answering a comprehension question correctly, indicating that these predictors partially account for the variability in response accuracy between participants. Comprehension questions were the largest source of group-level variability in overall response accuracy, while texts were the smallest source of group-level variability. The estimated negative covariance between participant-related variability in answering a comprehension question correctly and participant-related variability in the effect of perceived comprehension was again associated with considerable uncertainty (see Table 3.6).

3.3.3 Sensitivity Analysis

A sensitivity analysis was conducted to explore the robustness of the findings of the main analysis (section 3.3.2) to alternative analytical choices. In this section, the motivation for each of the alternative choices considered and how these alternatives were implemented is first discussed. Following this, the impact of the alternative model specifications on the estimated effects of interest are evaluated and a comparison of the predictive accuracy of the models is presented.

Alternative Model Specifications. Five analytical choices were identified as potentially impacting the findings of Study 1: i) the assumed shape of the predictive association between perceived and assessed comprehension, ii) the impact of multiple-choice questions options on the question response process, iii) the chosen wording of the judgement prompt, iv) the selection of priors for the effects of interest, and v) the use of the participant exclusion criteria. Each of these are discussed in depth below.

Nonlinear Relationship. The models defined in (7) and (8) assumed that the relationship between perceived and assessed comprehension is linear on the logit scale and that judgements of perceived comprehension can be treated as interval-level data. However, Nelson (1984) argues that there is no foundation for such assumptions. Given that judgements of perceived comprehension are ordinal, it may be the case that the difference between the units on the rating scale are not equivalent. For example, judging a text to be understood 'very well' may not correspond to double the level of perceived comprehension captured when judging a text to be understood 'slightly well'. In this situation, assuming linearity could produce estimates of the relationship between perceived and assessed comprehension which are misleading.

To examine the shape of the relationship between perceived and assessed comprehension, models which assumed a monotonic relationship were fit to the data. Monotonic effects refer to effects which remain in the same direction but differ in magnitude across the range of the variable. The monotonic effect of perceived comprehension was estimated using brms, following the approach described by Bürkner and Charpentier (2020). The parameters estimated for a monotonic relationship are not directly comparable to those for linear effects. Instead, the monotonic models estimate n-1 simplex parameters, where n is the total units on the rating scale, for the expected difference between adjacent ratings of comprehension and the overall average expected difference in probability of observing a correct response between ratings.

The models defined in (7) and (8) were altered to estimate the monotonic effect of perceived comprehension and associated interactions (Bürkner and Charpentier, 2020). Three simplex parameters were estimated in each monotonic model, corresponding to the change in the log odds of correctly answering a comprehension question given an increase in perceived comprehension from the second-to-third, third-to-fourth, and fourth-to-fifth judgements. Since there were no observations of the lowest response on the judgement scale, the change from first-to-second judgement could not be estimated.

Random Guessing. In the metacomprehension task, each of the four comprehension questions per text were presented with multiple response options. Previous research has shown that, when tasks with closed-form response options are employed, estimates of metacognitive ability become increasingly negatively biased (underestimated) as performance ability on the comprehension task decreases (Vuorre & Metcalfe, 2022). In multiple-choice tasks, individuals who are less able to correctly answer the questions correctly will rely on guessing to a greater extent than more able individuals. For less able participants, therefore, the predictive relationship between judgements of comprehension and observed performance may be undermined by correct guess responses. As a result, the population-level relationship between perceived and assessed comprehension may be underestimated in analyses which do not consider guessing.

One approach to accommodate correct responses which may be attributed to guessing is to assume a random guessing process occurs when participants know nothing on the performance task. This assumption can be incorporated statistically using a non-linear model in which the intercept is constrained with respect to the number of response options available (Bürkner, 2022). This approach models the effect covariates have on influencing the probability of correctly answering a comprehension correctly, in addition to some fixed probability of a correctly guessed response.

Given that each comprehension question in the metacomprehension task had four response options, if a response process of random guessing was used, a participant would have a 0.25 probability of selecting the correct answer by chance alone. Constraining the intercept to account for this probability, the model can be defined as:

$$Y_{ijk} \sim \text{Bernoulli}(P_{ijk}) \,,$$

$$P_{ijk} = \; 0.25 + 0.75 \log\left(\frac{\eta_{ijk}}{1 - \eta_{ijk}}\right),$$

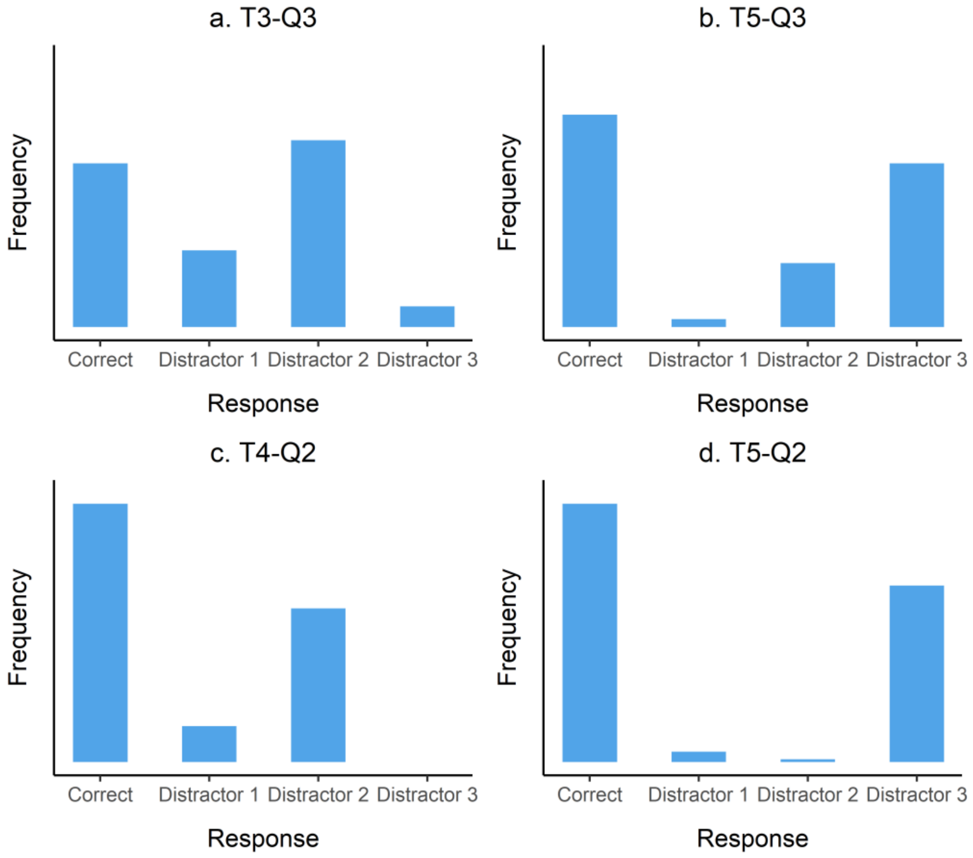
where the parameter η_{ijk} is the linear combination of the variables defined in models (7) and (8).

While this model offers an effective method to model probability of correct responses which can be attributed to chance alone, the assumption that random guessing occurs may not accurately describe the response process of participants (Higham, 2007; Vuorre & Metcalfe, 2022). Empirical support for random guessing in Study 1 can be evaluated by considering responses to comprehension questions with the lowest observed response accuracy. Contrary to an assumption of random guessing, Figure 3.7 shows that the selection of incorrect response options on questions where participants may be expected to engage in guessing does not appear random. Such differences in observed response selection may be due to participants typically having some knowledge relevant to the question, permitting the logical elimination of distractors (e.g., Embretson & Weztel , 1987), rather than relying on randomly selecting from the response alternatives.

However, while observed response selection appears incongruent with a random guessing process, this does not exclude the possibility that some element of guessing occurs in participants' response processes. Exploring an assumption of random guessing statistically provides a method to evaluate estimation bias that may be introduced if this assumption was true but not accounted for. Therefore, the models defined in (7) and (8) were altered to incorporate a random guessing response process.

Figure 3.7

Responses to Four Comprehension Questions with the Lowest Response Accuracy in Study 1



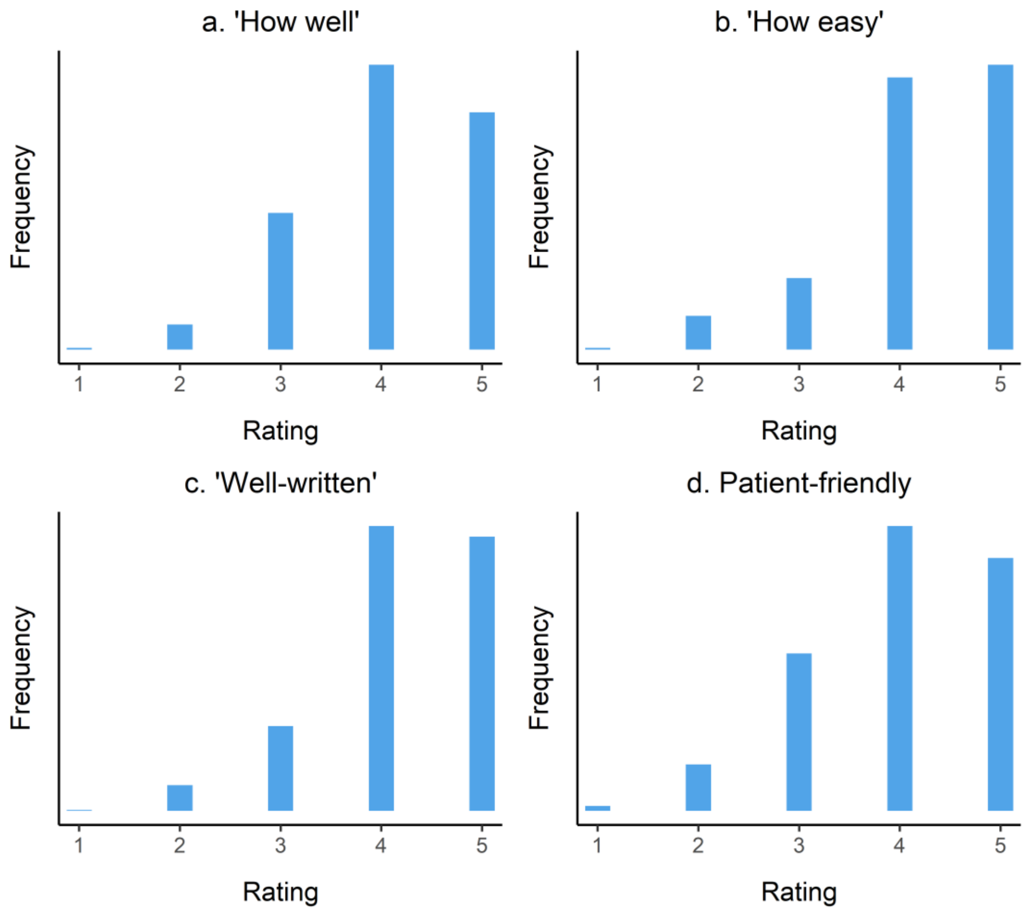
Note. T = text. Q = question.

Alternative Prompts. Given the research (Pilegard & Mayer, 2015), it was considered that the results of the main analysis may be sensitive to the wording of the judgement prompt. In the main analysis, the prompt ‘How much of the text do you feel you understand?’ was chosen to provide a measure of perceived comprehension. To permit comparisons between prompt wordings, judgements were collected using four other prompts on each text. Of these, two prompts elicited alternative measures of perceived comprehension and two prompts elicited judgements similar to those made by reader panellists on other aspects of the text.

The distributions of the rating responses to these prompts, shown in Figure 3.8, were broadly similar to the responses to the prompt used in the main analysis.

Figure 3.8

Distributions of Judgements in Response to the Four Alternative Prompts in Study 1



Note. ‘How well’ = ‘Overall, how well do you understand the text?’. ‘How easy’ = ‘How easy was it to understand the text?’. ‘Well-written’ = ‘How well was the text written? (In terms of spelling, punctuation and grammar)’. ‘Patient-friendly’ = ‘How patient-friendly was the text? (In terms of technical or medical words)’.

To examine the possible influence of prompt variation on the predictive relationship between judgements and assessed comprehension, the models defined in (7) and (8) were refit using observed responses to the alternative judgement prompts as the predictor

$Judgement_{ij}$. Alternative prompt judgements were standardised prior to model fitted and the linear relationship with assessed comprehension was estimated.

Variation in Priors. Weakly informative prior distributions were chosen for model parameters in the main analysis (section 3.3.2). Priors were selected which placed probability density diffusely over the plausible range of values for each parameter. Given the role of prior distributions in the resulting posterior distributions, it is important to assess the dependence of the effect estimates on the prior specification (Kruschke, 2015). If posterior distributions are not robust to reasonable variation in prior specifications, alteration to priors may considerably influence effect estimates and the resulting conclusions.

To evaluate the robustness of the posterior distributions, the models defined in (7) and (8) were refit using tighter (more precise) and looser (more diffuse) prior distributions for the effect parameters. The looser prior, specified as normally distributed with mean 0 and standard deviation 100, reflected comparably less a priori belief in effects of zero and greater belief in more extreme effect values. The tighter prior, specified as normally distributed with mean 0 and standard deviation of 1, reflected greater a priori belief in effects of zero or which are very small in magnitude.

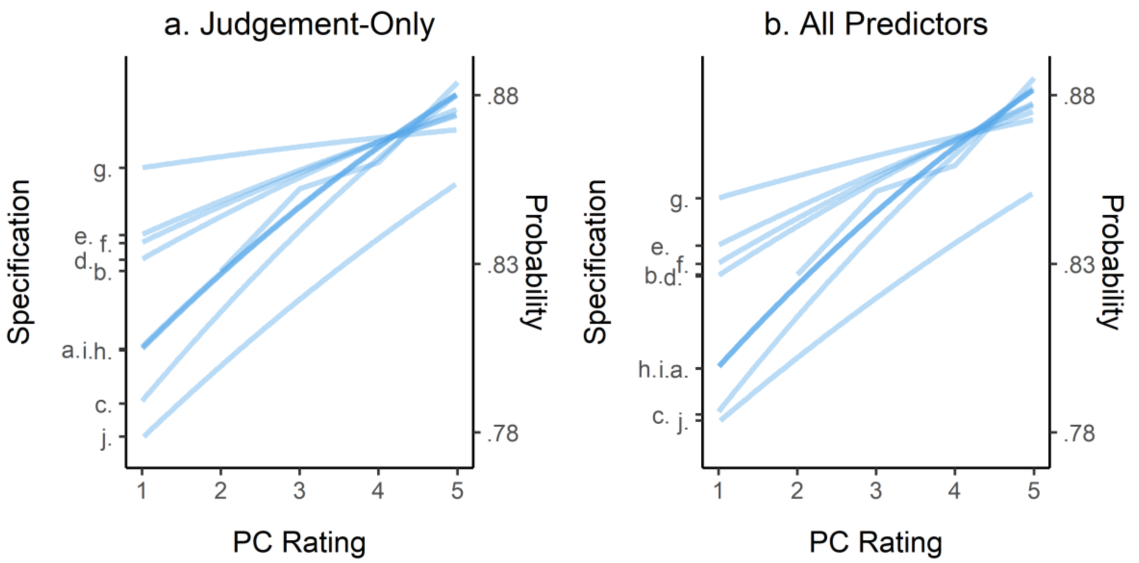
No Exclusions. Participants were excluded from the main analysis if any of their reading times for the 10 health-related texts were less than 30 seconds. The motivation for exclusion was to attempt to ensure validity in the measures of perceived and assessed comprehension. If participants had not read the text prior to making a judgement of their comprehension or answering a comprehension question, it is intuitive that such responses cannot be based on direct experience of reading the text. Therefore, to attempt to ensure reading has taken place, the amount of time spent with the text present on screen was used as a proxy of reading time.

In applying the exclusion criteria, the continuous measure of time spent on each text was dichotomised as < 30 seconds and ≥ 30 seconds. The practice of dichotomising continuous variables is recognised as problematic (Altman & Royston, 2006; Fernandes et al., 2019). The selected threshold also does not guarantee that an appropriate amount of reading has taken place to provide validity to judgements of perceived comprehension nor ensure individuals who are not engaging with the text are eliminated. Moreover, the findings from research which makes use of exclusion criteria may be dependent on the application of such criteria in unexpected ways, limiting the generalisability of the findings (e.g., Humphreys & Weisner, 2000). Given these concerns, the potential influence of the exclusion criteria on the results of the analysis was considered by fitting the models defined in (7) and (8) to the full dataset without exclusions (I = 245).

Model Comparison: Effects of Interest. For models which included only metacomprehension judgements as a population-level predictor, as in (7), across the alternative model specifications considered, estimates were broadly similar. The posterior means for the effect of judgements on the probability of correctly answering a comprehension question remained positive in all specifications. Further, limited participant-level variability in the predictive relationship between judgements and assessed comprehension was consistently found, with no negative participant-level model-fitted predictions of the effect estimated from any specification. The model-fitted predictions of the expected probability of answering a comprehension question correctly are shown in Figure 3.9a, for model specifications which include only metacomprehension judgements as a population-level predictor. Note that the range of the vertical axis is reduced in this figure to more clearly show the differences between the estimated effects.

Figure 3.9

Estimated Effect of Perceived Comprehension for Each Model Specification in Study 1



Note. Marginal model-fitted predictions at the population-level for model specifications including (a.) only participants' judgements as the predictor and (b.) all predictors are plotted. The probability of observing a judgement of '1' are a projection assuming a linear effect of perceived comprehension, since no participant gave this rating. This projection is not shown for the monotonic model. Model specifications a-j refer to: a. main analysis model, b. monotonic effect, c. constrained intercept, d. prompt 'Overall, how well do you understand the text?', e. prompt 'How easy was it to understand the text?', f. prompt 'How well was the text written? (In terms of spelling, punctuation and grammar)', g. prompt 'How patient-friendly was the text? (In terms of technical or medical words)', h. looser effect priors, i. tighter effect priors, j. no exclusion criteria. PC = perceived comprehension.

The greatest similarity to the estimated effect of metacomprehension judgements obtained from model (7) were observed in models which used looser and tighter prior distributions for population-level effects (labelled as specifications h. and i. in Figure 3.9, respectively). Similarly, despite the lower estimated intercept, when no exclusion criteria were applied (labelled as specification j.), the effect of metacomprehension judgements was equivalent to that estimated in the main analysis (labelled as specification a.). A lower intercept was similarly estimated for the model which incorporated a random guessing response process (labelled as specification c.), with a slightly larger effect of judgements estimated for this model. Assuming a random guessing response process, based on model-fitted predictions of the effect, a perceived comprehension judgement of understanding ‘none at all’ produces an expected probability of a correct response of 0.79 (*SE* = 0.05). In contrast, given a metacomprehension judgement of understanding 'all of it', the expected probability of a correct response increases to 0.88 (*SE* = 0.03). In addition, the estimated monotonicity in the relationship between perceived and assessed comprehension was limited (labelled as specification b.), indicated by the small deviations from the linear slope estimated in the main analysis.

The largest difference in the estimated effect of metacomprehension judgements, compared to the main analysis, was found for model specifications which used alternative judgement prompts as the population-level predictor. Principally, responses to the judgement prompt 'How patient-friendly was the text? (In terms of technical or medical words)' showed the weakest predictive relationship with assessed comprehension (labelled as specification g. in Figure 3.9). Based on model-fitted predictions of the effect, the expected change in the probability of answering a comprehension question correctly, given a judgement of ‘not at all’ to ‘extremely’, was an increase from 0.86 (*SE* = 0.04) to 0.87 (*SE* = 0.03). For the three other alternative prompts (labelled as specifications d., e., and f.), the estimated relationship between judgements and assessed comprehension was approximately half of the magnitude of effect reported in the main analysis. For all alternative judgement prompts, the 95% credible intervals spanned zero.

For models which included metacomprehension judgements, QRI scores, HLVA scores, and the interactions between these variables as population-level predictors, as in (8), estimates for the main effects were broadly similar across the alternative model specifications considered. Variation in the estimated relationship between perceived and assessed comprehension across model specifications was equivalent to that described above. This similarity can be seen in the model-fitted predictions of the expected probability of answering a comprehension question correctly are shown in Figure 3.9b. Estimated posterior means for the effect of judgements on the probability of answering a comprehension question correctly remained positive, with no negative participant-level model-fitted predictions of the effect, across alternative specifications.

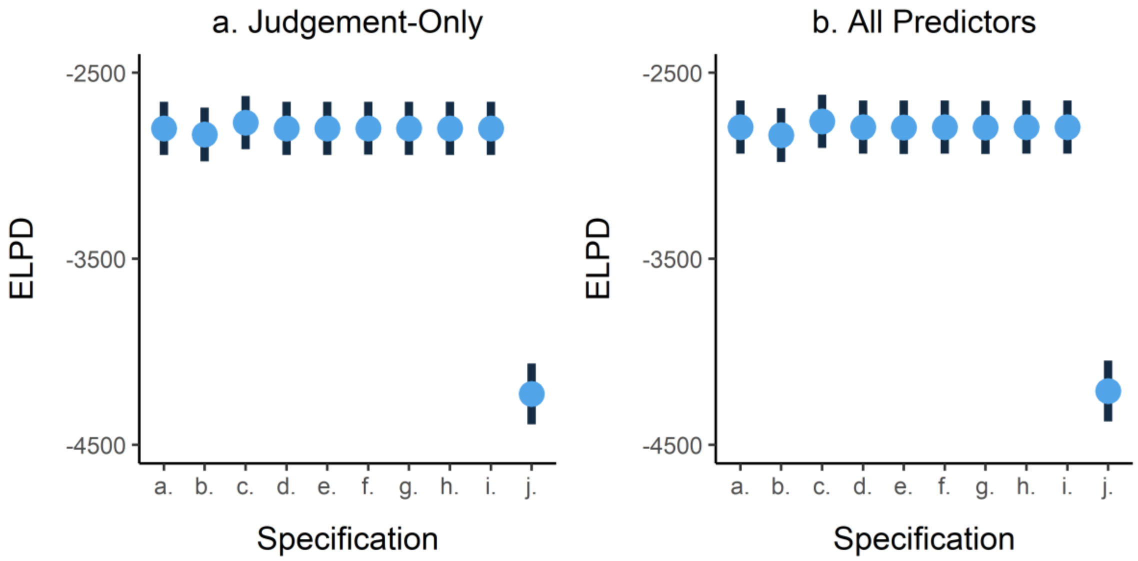
The main effects of QRI score and HLVA score were consistently estimated to be positive across alternative model specifications. Incorporating a random guessing response process and not applying the exclusion criteria resulted in slightly higher magnitudes for these effects. In contrast, with respect to the interactions between these variables and perceived comprehension, greater variability in the magnitude and direction was estimated across alternative specifications. When the exclusion criterion was not applied, support for the interaction between QRI score and perceived comprehension decreased, while support for a positive interaction effect between HLVA score and perceived comprehension increased. Given alternative judgement prompts 'How easy was it to understand the text?' and 'How patient-friendly was the text? (In terms of technical or medical words)', however, the posterior distributions for both interactions shifted towards negative values. This indicated less support for an interaction between QRI score and perceived comprehension and, concomitantly, greater support for a negative relationship between HLVA score and perceived comprehension in these models. The 95% confidence intervals for these effects nonetheless included both positive and negative values.

Model Comparison: Predictive Accuracy. A comparison of the relative predictive accuracy of the various model specifications was conducted to evaluate whether predictive performance differed across the models. Performance is often operationalised as a measure of the discrepancy between model predictions of the outcome and the actual observed data. Two methods were used to estimate the predictive accuracy of the models: i) calculating the estimated expected log predictive density (ELPD) and ii) posterior predictive checks. The ELPD was calculated, using the Widely Applicable Information Criterion (WAIC) computation (Vehtari et al., 2016; Watanabe, 2010), to provide a measure of out-of-sample predictive accuracy of each model. An ELPD closer to zero corresponds to a model which provides a better fit to the data than a model with larger, negative ELPD. Posterior predictive accuracy was evaluated by calculating the discrepancy between observed response accuracy and model predictions, to provide a measure of within-sample predictive accuracy (Gabry et al., 2019). Models which produce simulated frequencies of response accuracy which are comparable to the observed level of response accuracy provide a better fit to the data than models which do not.

With respect to out-of-sample predictive accuracy, Figure 3.10 shows the estimated ELPDs for each model specification, containing only metacomprehension judgements as the predictor of comprehension response accuracy (3.10a) and metacomprehension judgements, QRI and HLVA scores, and the interactions between these variables (3.10b). In this figure, the points show the estimated ELPD for each model and the vertical bars show the estimated uncertainty (three times the standard error of the estimated ELPD). Note that direct comparisons between ELPDs estimated from models fitted to different sized datasets (here, specification j., referring to the models fitted with no exclusion criteria) are not appropriate, given that possible differences in predictive accuracy are indistinguishable from the impact of additional observations on the calculation of the ELPD. Across the specifications fitted to the same sized datasets (a. to i.), differences in the estimated ELPD were minimal and were located within the intervals defined by three times the standard errors of the ELPD estimates, indicating highly similar out-of-sample predictive accuracy

Figure 3.10

ELPD Estimates for Each Model Specification in Study 1

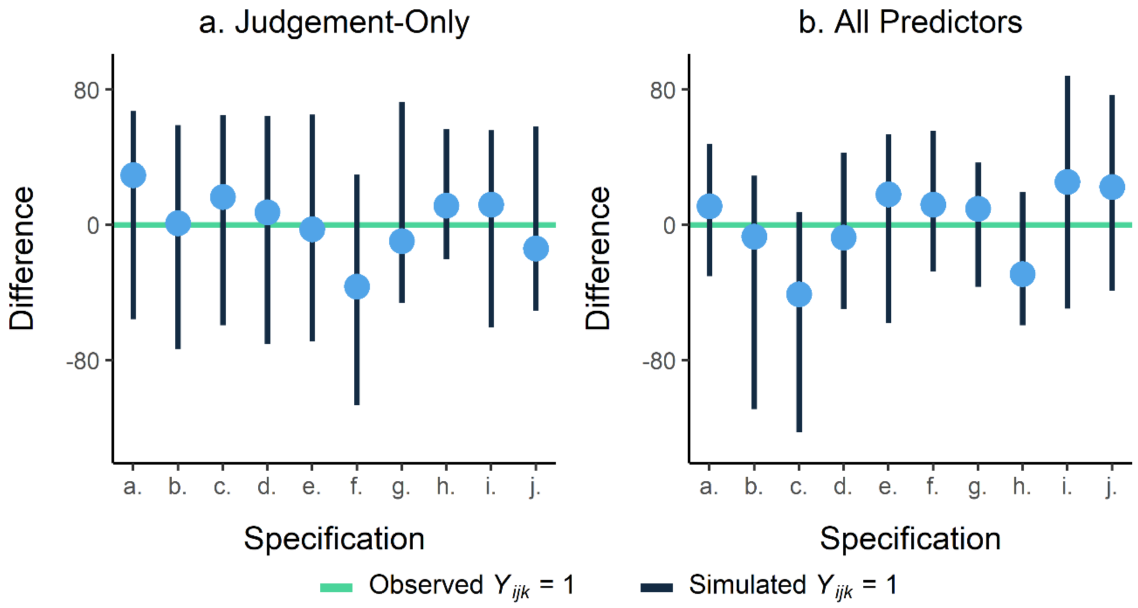


Note. Points (blue circles) show the estimated ELPD and bars show the standard error of the estimate, for model specifications including (a.) only participants' judgements as the predictor and (b.) all predictors. Model specifications a-j refer to: a. main analysis model, b. monotonic effect, c. constrained intercept, d. prompt 'Overall, how well do you understand the text?', e. prompt 'How easy was it to understand the text?', f. prompt 'How well was the text written? (In terms of spelling, punctuation and grammar)', g. prompt 'How patient-friendly was the text? (In terms of technical or medical words)'', h. looser effect priors, i. tighter effect priors, j. no exclusion criteria. ELPD = expected log predictive density.

With respect to within-sample predictive accuracy, Figure 3.11 shows the estimated posterior predictive accuracy for each model specification, containing either only metacomprehension judgements as the predictor of comprehension response accuracy (3.11a), or participants' judgements, QRI and HLVA scores, and the interactions between these variables (3.11b). In this figure, the light blue circles show the mean difference between observed and simulated correct responses and the dark blue lines show the range of differences estimated. Overall, across the model specifications, the discrepancy between observed and simulated response accuracy appeared limited (demonstrated by a mean difference which was located close to zero) and relatively unbiased (demonstrated by a range of estimated differences which was not located exclusively above or below zero).

Figure 3.11

Posterior Predictive Accuracy for Each Model Specification in Study 1



Note. Plotted according to model specifications including (a.) only participants' judgements as the predictor and (b.) all predictors. Light blue circles show the mean difference between observed correct responses ($Y_{ijk} = 1$) and simulated correct responses. Dark blue lines show the range of differences estimated between observed and simulated correct responses. Model specifications a-j refer to: a. main analysis model, b. monotonic effect, c. constrained intercept, d. prompt 'Overall, how well do you understand the text?', e. prompt 'How easy was it to understand the text?', f. prompt 'How well was the text written? (In terms of spelling, punctuation and grammar)', g. prompt 'How patient-friendly was the text? (In terms of technical or medical words)'', h. looser effect priors, i. tighter effect priors, j. no exclusion criteria.

3.4 Discussion

Study 1 was designed to address RQ1 and RQ2 (see section 3.1.1). In this section, the findings are discussed with respect to existing research, followed by a consideration of the theoretical implications of the results and the limitations of this research.

3.4.1 Main Findings

A weakly positive relationship between perceived and assessed comprehension was found in Study 1. This effect was robust in the sensitivity analyses, though the strength of the estimated relationship was lower for alternative prompts. Overall, these results indicate that judgements of perceived comprehension are weakly predictive of assessed comprehension. The main results indicate that, in the present sample of participants, texts and questions, given a judgement that none of the text is understood, participants are estimated to have approximately an 81% chance of successfully answering a comprehension question correctly, on average, despite this judgement implying little to no comprehension. In contrast, given a judgement that all of the text is understood, implying very high or complete text comprehension, the chance of answering a comprehension question correctly only increases to approximately 88%. The findings of Study 1, therefore, suggest that readers can typically be expected to demonstrate limited predictive accuracy in their metacomprehension judgements of health-related texts, with differences in these judgements corresponding to texts which are slightly more, or less, well understood.

In addition, limited variability between participants in the effect of perceived comprehension was found in Study 1. Limited variability in the predictive relationship was also replicated across the alternative analytical decisions explored in the sensitivity analysis. Model-fitted estimates suggest that individuals show an association between measures of perceived and assessed comprehension which is only slightly stronger or weaker than the overall average estimated relationship. This indicates that the capacity to provide metacomprehension judgements which are weakly predictive of understanding on health-related texts is on average shared across individuals.

As discussed in Chapter 2, previous research has primarily made use of correlation analyses to explore metacomprehension accuracy. Putting aside the limitations of such approaches discussed previously, studies which have adopted this method have demonstrated a weakly positive association between measures of perceived and assessed comprehension (Dunlosky & Lipko, 2007; Maki, 1998). The results obtained in Study 1, showing a weakly positive population-level effect of perceived comprehension on the probability of answering a comprehension question correctly, accords with the reported average correlation coefficients between perceived and assessed comprehension (Dunlosky & Lipko, 2007; Maki, 1998). Therefore, the present study provides support for the argument that there is a weakly positive correspondence between measures of perceived and assessed comprehension.

Contrary to previous research, however, individual variability in the relationship between perceived and assessed comprehension was found to be limited. Previous research involving correlation analyses has indicated individuals can show remarkably high, or even negative, associations between perceived and assessed comprehension (Chiang et al., 2010; Glenberg & Epstein, 1985; Jee et al., 2006). Variation on this scale between individuals was not observed in the present study: individuals varied marginally around a weakly positive effect. This finding suggests that the capacity to provide metacomprehension judgements which show some degree of predictive accuracy is not limited to particular individuals. This divergence in findings may be caused by differences in analytic approach. In contrast to the analyses reported in Study 1, studies which do not pool information in calculating individual-level estimates (i.e., 'no pooling' in estimating individual correlational measures of metacomprehension accuracy) can overstate the variation between individuals (Gelman & Hill, 2007). Given that Study 1 is the first to estimate variability in the predictive accuracy of metacomprehension judgements in this way (i.e., 'partial pooling' through hierarchical modelling), a replication study adopting the same analytic approach is required to verify this finding.

The analysis of how individual differences in reading ability and background knowledge may influence the relationship between perceived and assessed comprehension did not reveal any clear support for interactions between these variables and judgements of comprehension. Furthermore, the sensitivity analysis indicated that evidence in support of such interactions may be sensitive to analytic choices. The use of the exclusion criteria and the selection of the judgement prompt were both features which produced notable differences in interaction estimates. In the majority of models considered, however, a considerable portion of the posterior distribution for the interaction between reading ability and perceived comprehension judgements was located over values greater than zero, providing limited evidence that there may be a small positive relationship between these variables. The strongest support for an interaction between reading ability and metacomprehension judgements was observed when the model intercept was constrained to incorporate a random guessing response process.

With respect to reading ability, previous research regarding the relationship between metacomprehension accuracy and this variable has provided mixed findings. It has been suggested that there may be no relationship between metacomprehension accuracy and reading ability (Lin et al., 2000; Maki et al., 2005), or that a significant positive association between these variables exists (Griffin et al., 2008). The shifting evidence for a relationship between metacomprehension accuracy and reading ability, and the estimates obtained in Study 1, indicates that there may be a very small positive interaction between perceived comprehension judgements and reading ability. This would mean that individuals who score more highly on measures of reading ability provide judgements which are more predictive of their comprehension than individuals who score lower on measures of reading ability. Effectively, more able readers may be more likely to provide judgements which are more predictive of their comprehension outcomes. In explaining the difference in metacomprehension accuracy across reading ability, Griffin et al. (2008) contended that when processing demands are high, low ability readers may struggle to simultaneously engage in comprehension-building processes and monitor understanding at a meta-level. In contrast to the stimuli employed by Griffin et al. (2008), the texts used here were comparatively less challenging in terms of both difficulty (as measured by Flesch reading ease score) and length. It is therefore unlikely that the potential differences in the predictive accuracy of metacomprehension judgements emerged due to processing demands. Overall, however, given the previous findings and the uncertainty observed in the results of Study 1, the reliability of this interaction remains unclear.

Previous research has likewise provided conflicting evidence concerning the effect of background knowledge on metacomprehension accuracy, with the results being sensitive to how metacomprehension accuracy is operationalised. Research using relative (association-based) measures of metacomprehension accuracy has indicated that there may be no relationship between background knowledge and metacomprehension accuracy (Jee et al., 2006; Griffin et al., 2009; Shanks & Serra, 2014). In contrast, studies estimating absolute (magnitude-based) measures of metacomprehension accuracy have suggested that greater background knowledge may lead to predictions of performance which are closer to reality (Griffin et al., 2009; Jee et al., 2006) or may cause misplaced overconfidence in performance (Shanks & Serra, 2014). The estimates obtained in Study 1 indicate that variation in background knowledge does not influence an individual's capacity to demonstrate predictive accuracy in their comprehension judgements, consistent with research using association-based measures. As the difference in magnitude of judgements and performance was not considered in Study 1, it is not clear whether this independence is observed when this relationship is operationalised differently.

Although not a focus of the present study, the positive main effects of background knowledge and reading ability on the probability of answering a comprehension question correctly, observed in all models considered in the sensitivity analysis, is consistent with a wealth of comprehension research. The measure of reading ability captured a range of comprehension-based skills, for example, inferencing and reading strategies. The measure of background knowledge used in the present study was vocabulary-based, capturing domain-specific health-related knowledge via word knowledge. The sources of variance captured by the individual difference measures employed in Study 1 have been reported as predictive of comprehension outcomes and feature in multiple models of reading comprehension (Ahmed et al., 2016, Cromley & Azevedo, 2007; Gough & Tunmer, 1986; Perfetti, 1999; Perfetti & Stafura, 2014; Tzeng et al., 2005; van den Broek & Helder, 2017). The predictive utility of the HLVA and QRI in estimating the probability of answering a comprehension question correctly is therefore consistent with expectations based on prior research. Furthermore, the contribution of unit increases in both reading ability and background knowledge on the outcome was greater in magnitude than the effect found for perceived comprehension. This indicates that these variables are better predictors of the probability of answering a comprehension question correctly. However, perceived comprehension remained a reliable predictor following the inclusion of these variables in the model. It therefore appears that judgements of comprehension capture additional information which uniquely predicts variability in the probability of successfully demonstrating evidence of comprehension.

3.4.2 Theoretical Implications

Overall, the relationship between perceived and assessed comprehension evidenced in Study 1 suggests that individuals' metacomprehension judgements are weakly predictive of subsequent performance on health-related comprehension tasks. As discussed in Chapter 2 (section 2.1.1), since regression slopes are sensitive to differences in underlying discrimination accuracy (Rausch & Zehetleitner, 2017), it may be reasonable to consider that the measurement approach adopted in Study 1 provides some insight into underlying metacomprehension processes and that the predictive relationship is not produced merely as an artefact of measurement. As such, it is reasonable to consider the extent to which theoretical accounts of metacomprehension can accommodate the results observed in Study 1.

Consistent with the existing body of published research, a weak association between perceived and assessed comprehension was observed in Study 1 (Prinz et al., 2020a; Yang et al., 2022). In accounting for the weak relationship between metacomprehension judgements and performance on comprehension tasks, the dominant theoretical view holds that judgements are not informed by direct access to comprehension processes or products (Koriat, 1997). Direct access accounts suggest that metacognitive monitoring can take the internal representations created from processing stimuli as input in forming metacognitive judgements (Arbuckle & Cuddy, 1969; Hart, 1967), yet these accounts fail to explain why judgements are only weakly predictive of performance (Koriat, 1997; Schwartz, 1994; Tauber & Dunlosky, 2016). Instead, the limited association between perceived and assessed comprehension is argued to result from a metacomprehension judgement process which relies on using various sources of information (cues) in order to infer the quality of understanding gained from a text (Koriat, 1997; Griffin, Mielicki & Wiley, 2019).

According to cue-based accounts of metacomprehension judgements, individuals variously select from available cues to inform their judgements. The selection of cues is not well-specified within these accounts, but may be determined by salience, heuristics or knowledge-driven processes (Dunlosky, Rawson & Hacker, 2002; Koriat, 1997; Griffin et al., 2009; Griffin, Mielicki & Wiley, 2019; Linderholm et al., 2008; Zhao & Linderholm, 2008).

Nevertheless, these accounts suggest that, on average, individuals select cues which are poorly predictive of underlying comprehension, resulting in the observed weakly positive average association. The overall weak association occurs due to the unavailability of or the failure to select cues with high predictive validity (Griffin, Mielicki & Wiley, 2019). In addition, how effectively a cue can predict comprehension outcomes also varies with situational factors, such as the nature of the comprehension assessment (Dunlosky, Rawson & Hacker, 2002; Koriat, 1997; Wiley et al., 2005). Consequently, cue selection may interact with experimental factors, thereby contributing to the low average level of association (Griffin, Wiley & Thiede, 2019).

However, alongside the average weak association between perceived and assessed comprehension, considerable between-individual variability in metacomprehension accuracy has been reported. Previously, individuals have been estimated to vary considerably in their ability to accurately judge their level of understanding, with some individuals showing near perfect while others show a negative association between judgements and observed performance (Chiang et al., 2010; Glenberg & Epstein, 1985; Jee et al., 2006). These estimates mirror the marked individual variability in self-reported cue use which has previously been found (Jaeger & Wiley 2014; Thiede et al., 2010). According to cue-based accounts of metacomprehension judgements, the wide range of observed metacomprehension accuracy reflects the variability in cue use across individuals (Koriat, 1997; Thiede et al., 2010).

A cue-based account of metacomprehension accuracy may be a cogent explanation for the average weakly predictive relationship between measures of perceived and assessed comprehension observed in Study 1. Individuals may have selected cues which provided minimal diagnostic accuracy to form their judgements. However, the near equivalence between individuals in the predictive relationship between perceived and assessed comprehension is incongruent with cue-based accounts which suggest cue selection varies considerably between individuals. To accommodate the findings of Study 1 within such accounts, either i) participants selected a variety of cues which were similarly predictive of comprehension outcomes or ii) participants primarily relied on the same source of information in forming their metacomprehension judgements. Given that both of these aspects are assumed to vary within existing cue-based accounts, it appears to be challenging for these accounts to explain the lack of variability in the predictive relationship observed in Study 1.

Yet, within cue-based frameworks of metacomprehension, the use of a shared default cue has previously been discussed. For example, within the situation-model view of metacomprehension, in the context of informing participants of an upcoming test on the material, Griffin, Wiley and Thiede (2019) suggest that individuals default to a memory-based assessment heuristic. Similarly, within Dunlosky, Rawson & Hacker's (2002) levels-of-disruption account, it is suggested that processing disruptions during text comprehension are the primary cue which individuals use to form metacomprehension judgements. Further, in the context of an anchoring and adjustment model of metacomprehension accuracy (Zhao & Linderholm, 2008), Linderholm et al. (2008) suggested that individuals may select a common estimation point as the foundation for metacomprehension judgements, such as shared beliefs about the average population-level performance on a task.

Perhaps, in the context of a highly controlled experiment, a homogenous environment might plausibly provide a basis for homogenous cue use. Limited variability between individuals, in these circumstances, could reasonably be accommodated with cue-based accounts. However, the shared default cues which have previously been suggested are limited in their capacity to account for the homogeneity observed in Study 1 (Dunlosky, Rawson & Hacker, 2002; Griffin, Wiley & Thiede 2019; Linderholm et al., 2008). Firstly, it is unlikely participants defaulted to a memory-based assessment heuristic or relied on beliefs about average test performance in the wider population, given that participants were not explicitly informed of an upcoming test prior to providing metacomprehension judgements. Secondly, the potential shared default cue is also unlikely to be disruptions which occurred during reading, with such disruptions treated as equivalently predictive of comprehension. Since no disruptions were inserted into texts, experienced processing disruptions can be expected to have varied between individuals (Baker, 1979; Winograd & Johnston, 1980). As such, the severity of experienced disruptions, the level of text processing disrupted, and whether or not the disruption was resolved would likely influence how processing disruptions align with task performance.

While previously suggested default cues may not readily account for the findings of Study 1, the results reported here nevertheless indicate a default tendency to attend to a primary source of information. Rather than metacomprehension judgements variously reflecting task experiences or beliefs which are independent of the task across individuals, judgements may reflect a shared source of information. Such commonality in cue use would suggest that cue selection processes are fundamentally similar across individuals. Moreover, identifying the shared source of information would provide insight into the conditions which drive cue selection. Further, in the context of the current research, identifying the source of information would provide insight into the utility of reader panellists' judgements of health-related texts. As cue selection processes are underspecified in current theoretical accounts of metacomprehension judgements, further research would be beneficial in progressing our understanding of metacomprehension judgements.

3.4.3 Limitations

It is important to note two limitations of Study 1 which relate to observed variability. Firstly, the limited variability in the measures of perceived and assessed comprehension may curtail the generalisability of the reported results. Question performance was high overall, indicating that the majority of participants were able to answer most of the questions with ease. It may be the case that the findings here relate only to cases where the text is easily understandable, and where the majority of questions are readily answerable. In addition, ratings of comprehension were highly skewed, with the bulk of perceived understanding responses being 'most of it' and 'all of it'. The lack of observations at lower ratings of perceived comprehension (with no ratings of ‘none at all' given) means that it is harder to be confident about the relationship between perceived and assessed comprehension at these lower ratings. While there is sufficient variance in this sample of participants, texts and questions to satisfactorily describe the relationship between perceived and assessed comprehension observed in Study 1, this relationship may differ where there is more variability in perceived and assessed comprehension. Further research with a different sample of participants, texts and questions, as in the following study in the present research, would be beneficial in clarifying the specificity of the observed findings.

A second factor which may limit the findings of Study 1 relates to the estimated question-level variability on the probability of answering a comprehension question correctly. Questions contributed considerable variability in achieving the outcome, indicating that question difficulty was not homogenous. This variability may arise for several reasons, such as the depth of required background knowledge to comprehend the proposition(s) the question assesses (Kintsch & van Dijk, 1978), the efficiency of distractor items (Ebel & Frisbie, 1991; Graesser et al., 2009), or whether the question assesses an automatic or more elaborative inference (Graesser et al., 1994). To date, studies examining metacomprehension accuracy have not explicitly estimated question-level variability, meaning the results reported in Study 1, while informative, are problematic in permitting comparisons with previous research. It would appear that, although the questions in Study 1 were not uniformly challenging, the distribution of difficulty was somewhat random across texts - each text contained a mixture of more or less difficult questions. This means that the correspondence between perceived and assessed comprehension was likely not undermined by systematic differences in question difficulty between texts. However, question-level differences are clearly a source of variance which can influence measures of comprehension, and therefore estimates of the predictive relationship between perceived and assessed comprehension (Wiley et al., 2005). This issue is considered in subsequent studies in the present research.

3.4.4 Conclusion

Study 1 found a weakly positive predictive relationship between perceived and assessed comprehension, with limited individual variability in the strength of this relationship. Individual differences in background knowledge and reading ability did not reliably influence the relationship between perceived comprehension judgement and assessed comprehension. Judgements of comprehension, therefore, provide some information of predictive validity which is useful in improving the likely comprehensibility of health-related information. Moreover, given the limited variability in the informativeness of judgements on comprehension outcomes across individuals, it is likely that the judgements of individuals on reader panels are similarly predictive of comprehension.

To the extent which the estimated relationship may inform of processes occurring during the formation of metacognitive judgements, a cue-based theory of metacomprehension may account for the limited correspondence. However, existing explanations fail to satisfactorily account for the lack observed variability in the predictive validity of participants' judgements. An alternative view, that metacomprehension judgements are based on a primary shared source of information may more plausibly account for the findings of Study 1.

Further research is required to address a number of highlighted issues. Firstly, it is important to replicate the observed low individual variability in metacomprehension accuracy using the analytic approach taken here. Additionally, a new sample of participants, texts and questions is required to examine the generalisability of these findings. Furthermore, to progress our understanding of cue selection processes, as considered in the following study in the present research, further work is required to consider whether metacomprehension judgements are based on a shared source of information across individuals and what this primary cue may be.

4. Study 2

In this chapter, the motivation for the second study (Study 2) is discussed, leading to the identification of two research questions to be addressed. In the method section, firstly, the identification of the required sample size is considered in a design analysis. Following this, a pilot study and the participants, materials, and procedure of Study 2 are described. The presentation of results is separated into a preliminary data inspection, the main planned analysis, and an analysis of the sensitivity of the main findings to alternative analytical choices. Lastly, the findings of Study 2 are discussed.

4.1 Introduction

The apparent homogeneity in the predictive relationship between individuals observed in Study 1 is poorly accommodated within theoretical accounts of metacomprehension that suggest that the usage and predictive validity of cues varies considerably. As discussed in Chapter 3 (section 3.4.2), the results of Study 1 may instead be more plausibly accounted for by a default tendency to rely on a shared source of information across individuals. Primary judgement cues which have previously been suggested arguably do not fully account for the observed results (Dunlosky, Rawson & Hacker, 2002; Griffin, Wiley and Thiede, 2019; Linderholm et al., 2008). Identifying the primary cue used to inform metacomprehension judgements, should it exist, would provide insight into what reader panellists' judgements predominantly indicate and, therefore, whether they are likely to be useful in evaluating health-related texts. In pursuit of better understanding reader judgements, the remainder of this section explores one potential source of information which may plausibly be proposed as underlying metacomprehension judgements.

Drawing on the levels-of-disruption hypothesis and the concept of hierarchical semantic structures, discussed in theories of text comprehension (e.g., Kintsch, 1988), metacognitive experiences relating to the construction of a coherent macropropositional structure of the text may be identified as a potential primary source of information in metacomprehension judgements. To explore this possibility, theoretical and empirical research concerning hierarchical semantic structures within text comprehension is first discussed in the following section. Subsequently, evidence relating to how individuals may metacognitively monitor information varying in semantic importance within text is considered, augmenting the results of Study 1. Finally, complexities in experimentally exploring the role of semantic importance in metacomprehension judgements are briefly discussed, before moving on to the specific research aims of Study 2.

4.1.1 Metacomprehension Judgements and Macropropositional Coherence

The concept of semantic structures is embedded in multiple theories of text comprehension, with several accounts suggesting that the network of meanings and ideas, arising from an episode of reading, form a relational-hierarchy of importance or feature more or less semantically connected elements (Graesser et al., 1994; Johnston & Afflebacher, 1982; Kintsch, 1988; Mandler & Johnson, 1977; Thorndyke, 1977; Trabasso & Sperry, 1985; Trabasso & van den Broek, 1985; van Dijk & Kintsch, 1983; Zwaan & Radvansky, 1998). In Kintsch's (1988) construction-integration framework, for example, semantic hierarchies form a key element of text processing, with the comprehension of texts strongly influenced by a schematic superstructure which defines what information is considered important (Kintsch, 1994; Kintsch & van Dijk, 1978; van Dijk & Kintsch, 1983). Specifically, during reading, information is first parsed into micropropositions, which are then converted into macropropositions to 'fill the slots' in the schematic superstructure (Kintsch, 1994; Kintsch & van Dijk, 1978; van Dijk & Kintsch, 1983). The resulting macropropositions are hierarchically structured, with some macropropositions featuring at higher, or multiple levels of importance or relevance within the situation model. Macropropositions can be regarded as representing the gist of the text and, at higher levels of abstraction, represent the topic (Kintsch, 1994; Kintsch & van Dijk, 1978; van Dijk & Kintsch, 1983).

In support of the concept of hierarchically structured text representations, evidence of the impact of semantic and structural importance, or ‘centrality’, has previously been observed. The ability to identify and recall information which is central to the semantic structure of the text, even where this requires some level of abstraction, appears to emerge early in development and improve with age (Brown & Smiley, 1977; Lynch et al., 2008; Otto et al., 1969; van den Broek et al., 1996). Developing and adult readers recall more information which is considered central to the semantic structure of the text after reading than elements considered less central, in both free recall and summary-writing tasks (Kim et al., 2008; Kintsch & van Dijk, 1978; Omanson, 1980; McCrudden et al., 2011; Thorndyke, 1977; Trabasso et al., 1984; van den Broek & Trabasso, 1986; Yeari et al., 2015; Yeari et al., 2017). In addition, reading time data show that skilled readers spend more time processing central information, compared to less central information, during reading (Yeari et al., 2015; Yeari et al., 2017). Research has also indicated that word and sentence-level semantic and integration processes during reading can be influenced by centrality (Helder et al., 2019; Nieuwland & Van Berkum, 2006; Stafura & Perfetti, 2014). Moreover, observed quantitative and qualitative differences in the recall of central information has prompted suggestions that preferentially processing and attending to central information may be considered a skill itself (Kendeou et al., 2009; van den Broek et al., 2013; van den Broek & Helder, 2017; Yeari & Lantin, 2020).

Despite recognition that the importance of text elements varies in comprehension processes, there is notable variability in the conceptualisation of centrality. For example, views differ on whether centrality-related reading processes are top-down, bottom-up or a mix of the two (Helder et al., 2019; Kintsch & van Dijk, 1978; Mandler & Johnson, 1977;

Omanson, 1980; Thorndyke, 1977; van den Broek et al., 2013; van den Broek & Helder, 2017; van Dijk & Kintsch, 1983). There is also debate concerning whether centrality and relevancy are separable constructs. Yeari (2015) argues that centrality is determined by text structure and content, whereas relevancy is dependent on the reader's goals. In contrast, van Dijk and Kintsch (1983) suggest that the schematic superstructure, governing microproposition relevance and incorporation into the macrostructure, may be strongly influenced by the reader's goals and context, such that relevance determines centrality.

Variability in the conceptualisation of centrality has led to differences in how centrality is measured, with ambiguity in the variables underpinning the definitions of centrality presenting a challenge for effective measurement. For example, it is contended that central text elements can be identified as having comparably greater connections with multiple text elements than non-central elements (Albrecht & O'Brien, 1991; Helder et al., 2019; Miller et al., 2013; Trabasso & Sperry, 1985; van den Broek, 1988; van den Broek et al., 2013). Accordingly, measurement of centrality would require defining both what constitutes a text element and what constitutes a connection. Previously, text elements have been operationalised as micro-propositional level units (i.e., Kintsch, 1988) or have been defined more coarsely at the 'idea unit' level (e.g. Helder et al., 2019). Similarly, in defining a connection between information, researchers have variously considered the degree of semantic-relatedness (Mo et al., 2007), the frequency of mentions in a passage (Albrecht & O'Brien, 1991, Helder et al., 2019), or the extent to which it would impair comprehension if omitted (Yeari et al., 2015).

More pragmatically, in order to measure centrality or to validate measures of centrality, researchers may collect ratings of perceived centrality. In support of this approach, rated centrality has been found to align with behavioural measures of text comprehension, such as recall and reading time (Yeari et al., 2015; Yeari et al., 2017). However, it is unclear what processes readers engage in when providing judgements of centrality or how these perceptions may correspond to the construct. Arguably, valid ratings are underpinned by adequate comprehension of the text: to extract the central ideas, readers need to successfully engage in comprehension-building processes and form an inter-connected representation of the meanings in the text (Kintsch & van Dijk, 1978; van den Broek et al., 2013). However, adequate comprehension is typically not considered alongside centrality ratings. Eliciting ratings of centrality may not, therefore, be an optimal way to address the challenges in objectively operationalising centrality.

Despite the variability in the definition and measurement of centrality, differing accounts are underpinned by the view that some elements are more semantically and structurally important in text comprehension (Helder et al., 2019; van den Broek et al., 2013). Further, elements which are central to the semantic structure of the mental representation can be described as capturing the main ideas of the text (Kintsch & van Dijk, 1978; Thorndyke, 1977; van den Broek & Helder, 2017; van Dijk & Kintsch, 1983). This has clear implications for understanding what it means to achieve comprehension of a text. For main ideas to be understood and reported following reading, readers must produce a well-structured and well-connected text representation (van Dijk & Kintsch, 1983; Yeari & Lantin, 2020). Obtaining a coherent set of macropropositions, therefore, is a product of engaging in successful comprehension-building processes. In addition to this, researchers have suggested that understanding a text in a way consistent with the author's intentions involves extracting and comprehending the main ideas that the author wishes to communicate (Kintsch, 1994; Kintsch & van Dijk, 1978). It can be argued, therefore, that understanding central elements, as intended by the author, simultaneously results from successful text comprehension and constitutes the desired comprehension outcomes (Johnston & Afflebacher, 1982; Yeari et al., 2017; Yeari & Lantin, 2020).

If metacomprehension judgements were informed by the success of establishing a coherent macropropositional structure, judgements would provide valid insight into the quality of text comprehension. For example, encountering unresolved disruptions to the comprehension of information which is semantically central would likely be an impediment to establishing a coherent semantic structure. As such, the capacity to construct macropropositions would be particularly informative of comprehension outcomes. Meaningful insight into text comprehension could, therefore, be provided by metacognitive experiences, generated through comprehension monitoring at the macropropositional level of text processing. Similar to suggestions made within both the situation-model-cues account and levels-of-disruption hypothesis (Dunlosky, Rawson & Hacker, 2002; Griffin, Wiley & Thiede, 2019), these metacognitive experiences may be used to form metacomprehension judgements which directly correspond to the status of an individual's underlying text comprehension. In contrast to these accounts, however, the coherence of information which is semantically central may be proposed as a primary cue, rather than other aspects of the situation model. Importantly, for such information to underlie metacomprehension judgements, comprehension monitoring processes must be sensitive to centrality.

Previous research which has examined whether comprehension of central elements may be preferentially monitored has provided mixed results. In support of the notion that comprehension monitoring is sensitive to semantic centrality, Yussen and Smith (1990) found that inconsistencies introduced in the overall meaning of multiple sentences were reported more frequently and lead to lower ratings of text comprehensibility than inconsistencies at the level of specific details within single sentences. Similarly, Baker (1979) observed that low error reporting rates were accompanied by participants claiming that they had understood the central idea of the text, suggesting that judgements of text consistency were influenced by whether sufficient overall text comprehension was achieved despite the inconsistency. In contrast to these findings, however, other studies have failed to demonstrate an effect of centrality on error detection, indicating that monitoring may not preferentially attend to macroproposition coherence. Baker and Anderson (1982) report that inconsistencies which relate to central and peripheral elements were identified with similar frequency.

One reason which may account for the variability in the relationship between centrality and error detection is that individuals vary in the desired depth of understanding, favouring the construction of a more or less rich macropropositional structure. This assertion is consistent with the standards of coherence framework, which suggests that the types and strengths of connections a reader may wish to make during reading are variable and are influenced by their purpose for reading purposes (Calloway, 2019; van den Broek et al., 2011; van den Broek & Helder, 2017). As a consequence of variability in the desired macropropositional richness, individuals will differ in which ideas are considered semantically central (Baker & Anderson, 1982; Brown & Smiley, 1977). For example, in Baker and Anderson's (1982) study, 9% of main-point and 41% specific-detail text inconsistencies were classified as the opposite type by participants. Observing similar rates of inconsistency detection for central and peripheral elements may, therefore, result from consistency manipulations impacting on macropropositional coherence in ways contrary to expectations.

While the desired richness of understanding is expected to vary by individual and context, readers may minimally attempt to obtain a coherent macropropositional structure. This structure may be poorly interconnected or represent a subset of the available macropropositions yet be sufficient to satisfy their purpose for reading (van den Broek et al., 2011). In effect, readers may aim to extract a coherent gist, adequate in capturing the elements which they consider central. Parallels between this suggestion and the 'good-enough' view of syntactic processing in text comprehension may be drawn to support this view (Christianson et al., 2001; Ferreira et al., 2001; Ferreira et al., 2002; Kintsch, 1988). Readers may frequently use fast and efficient processing to obtain sufficient meaning, with partial understanding, or complete omission, of some elements considered non-detrimental to overall understanding of the text (Christianson, 2016). Individuals may be content with their overall understanding, despite this being incongruent with the author's intended meaning (Christianson, 2016; Ferreira et al., 2002). Monitoring one's comprehension of central information may not, therefore, correspond to monitoring the coherence of the macropropositional structure as intended by the author.

While macropropositional structure may vary between individuals, it is reasonable to expect some similarity in the information captured within macropropositions, conditional on the goal of reading. The purpose of reading for understanding likely corresponds most closely to the author's expectations of a reader's behaviour (Kintsch, 1994). In this situation, reader-author variability in the macropropositional structure of the text is likely to be lower than given alternative purposes for reading. Furthermore, given the same purpose for reading, macropropositional structures may show greater similarity between individuals compared to situations in which purposes vary across readers (Kintsch & van Dijk, 1978; Lehman & Schraw, 2002; Narvaez et al., 1999). Assuming a shared goal of reading for understanding, therefore, it is more likely that a reader's set of constructed macropropositions overlaps both with other readers’ and the author’s.

The account proposed above, that metacomprehension judgements are informed by metacognitive experiences concerning the coherence of macropropositions, may explain the findings of Study 1. Notwithstanding the same purpose for reading, due to differences in reading ability, background knowledge and standards of coherence, macropropositions will vary on several dimensions, such as whether macropropositions contain accurate interpretations. Limited individual variability in the predictive relationship between perceived and assessed comprehension could reflect coherence monitoring of macropropositions that differ slightly between skilled readers. The overall weak predictive relationship, meanwhile, could be attributed a misalignment between the basis of metacomprehension judgements and the performance measure of comprehension (Dunlosky, Rawson & Hacker, 2002; Wiley et al., 2005). In Study 1, comprehension questions aimed to assess the capacity to generate inferences, in relation to information identified from gist-level summaries. Accordingly, if judgements are informed by a default attendance to macropropositional coherence, a stronger predictive relationship may be expected when questions directly assess comprehension of semantically central information compared to less central (or peripheral) information.

Investigating the potential relationship between centrality and metacomprehension, however, is complicated by the possibility of a reliance on the same comprehension-building processes shared across the semantic hierarchy of the text representation. Due to this dependence, comprehension of information which is more versus less semantically central will be likely to share variance. Therefore, we may expect that judgements based on the comprehension of semantically central information will also show a predictive association with the comprehension of semantically peripheral information. Importantly, the extent to which a reliable difference may be observed in the magnitude of the predictive relationship between perceived and assessed comprehension of central and peripheral information is conditional on the strength of the dependence between these two outcomes. Unfortunately, research has not previously considered the strength of this association, which may differ between readers and texts.

While it may be a challenge to observe evidence consistent with the claim that macropropositional coherence underlies metacomprehension judgements, experimentally exploring this would nonetheless provide insight into metacomprehension judgements. In the context of reader panels, such research would allow us to quantify the extent to which judgements inform of likely comprehension outcomes of information which is semantically central, peripheral or both. This could identify the aspects of health-related texts that metacomprehension judgements most closely associate with and, therefore, are most useful in evaluating. By investigating the potential sensitivity of metacomprehension judgements to semantic centrality, therefore, we may improve our understanding of how reader panel judgements can be used most effectively in the production of healthcare information.

4.1.2 Research Aims of Study 2

To explore the potential role of macropropositional coherence in metacomprehension judgements, Study 2 was designed to consider whether the predictive relationship between metacomprehension judgements and assessed comprehension may be influenced by the semantic centrality of information in the text. In addition, to examine whether the findings of Study 1 could be reproduced, measures of reading ability and background knowledge were included as potential moderators of the predictive relationship between perceived and assessed comprehension.

A third individual difference measure was also explored, given the role of standards of coherence in text comprehension (van den Broek et al., 2011). Since the types and strengths of coherence a reader seeks to establish during reading shapes the macropropositional structure, variability in standards of coherence may influence the predictive relationship between judgements and comprehension of semantically central information. To explore this possible interaction, the impact of individual variability in reader-based standards of coherence (Calloway, 2019) was also examined. Study 2, therefore, aimed to address two research questions:

> **RQ3:** How does the strength of the relationship between judgements of comprehension and assessed comprehension vary depending on the semantic centrality of the information assessed on health-related texts?

RQ4: How are individual differences in reading ability, background knowledge and standards of coherence related to variation in the relationship between perceived and assessed comprehension?

Potential limitations identified in Study 1 were also addressed alongside these research questions. Firstly, considerably negative skew in metacomprehension judgements was observed in Study 1, meaning that limited information was obtained for low judgements of perceived comprehension. As a result, greater uncertainty in the relationship between perceived and assessed comprehension was estimated across lower judgement ratings on the metacomprehension scale. To address this, more challenging health-related texts were included to increase the chance of observing a wider range of ratings. It was also anticipated that this would produce lower levels of overall accuracy in assessed measures of comprehension, addressing potential concerns that judgements may only be predictive on texts which are readily comprehended. In addition, the number of response options available for rating perceived comprehension was increased from a 5-point to a 7-point scale. This alteration was made based on the assumption that permitting finer gradations of comprehension judgements may produce a better range of rating responses. However, evidence in support of this view is mixed (Dawes, 2008; Leung, 2011; Preston & Colman, 2000).

4.2 Method

Ethical approval for this study was granted (for piloting and data collection) in July 2020. Piloting and data collection commenced in October 2020. A detailed preregistration for this study, including materials and analysis plan, was uploaded to an OSF repository in October 2020 (https://osf.io/2kasd/).

4.2.1 Design Analysis

To evaluate the capacity to estimate the effect of interest with accuracy and precision, an analysis of the proposed design for Study 2 was conducted under varying sample sizes. Achieving accuracy and precision in estimation was operationalised as described in Chapter 2 (section 2.3.1). The sample size of Study 2 is considered adequate when the probability of achieving the target level of accuracy and precision in estimation is expected to occur in at least 80% of hypothetical studies.

The two research questions for Study 2, RQ3 and RQ4, concern five parameters of interest: i) the effect of perceived comprehension for central information, ii) the effect of perceived comprehension for peripheral information, iii) the interaction between perceived comprehension and reading ability, iv) the interaction between perceived comprehension and background knowledge, and v) the interaction between perceived comprehension and standards of coherence. It was considered that Study 2 should primarily have the capacity to address RQ3 for two reasons. Firstly, as interactions between individual difference measures and metacomprehension judgements in the context semantic centrality have not previously been investigated, the presence and magnitude of these interactions was highly uncertain. Secondly, the level of precision for the effect of metacomprehension judgements was similar to the interaction effects in Study 1. Consequently, it may be reasonable to expect similar levels of accuracy and precision in estimation are observed both for both the effect of metacomprehension judgements and the interaction effects in Study 2.

Given the above, a multilevel logistic regression model (described below) was used to simulate response data to evaluate the capacity to estimate, with accuracy and precision, the slope coefficients for perceived comprehension of central information and for perceived comprehension of peripheral information. The selection of model parameter values was informed by Study 1 data and previous research, where appropriate.

Assessed Comprehension Simulation Model. In the simulations, the binary response *Y* denotes whether or not an individual participant *i*, reading a health-related text *j*, answered a comprehension question *k* correctly ($Y_{ijk} = 1$) or incorrectly ($Y_{ijk} = 0$). Letting P_{ijk} be the corresponding probability of observing a correct response: $P(Y_{ijk} = 1)$, this event can be expressed as the result of a Bernoulli trial:

$$Y_{ijk} \sim \text{Bernoulli}\left(P_{ijk}\right),$$

where *i* = 1, …, *I*, *j* = 1, …, *J* and *k* = 1, …, *K*; and *I*, *J* and *K* refer to the total number of participants, texts and questions per text, respectively. Assuming a logit-link function, the explanatory variables of interest and hierarchical sources of variance were linked to the response:

$$\log\left(\frac{P_{ijk}}{1 - P_{ijk}}\right) = \beta_0 + (\beta_1 + u_{1i})Judgement_{ij} \ + (\beta_2 + u_{2i})Centrality_k \qquad (9)$$

$$+(\beta_3 + u_{3i})Judgement_{ij}Centrality_k + u_{0i} + v_j + w_k,$$

where $Judgement_{ij}$ refers to the observed rating of perceived comprehension from individual *i* in response to text *j*; $Centrality_k$ refers to the observation of whether question *k* assesses semantically central or peripheral information; β_0 refers to the intercept (baseline log odds of observing success); β_1 refers to the population-level effect of a perceived comprehension judgement for semantically central information; β_2 refers to the population-level effect of semantic centrality; β_3 refers to the population-level change (interaction) in the effect of a perceived comprehension judgement given semantically peripheral information; u_{1i}, u_{2i} and u_{3i} refer to individual-level variability in the respective population-level effects; and u_{0i}, v_j and w_k refer to intercept variability at the level of the individual, text and question, respectively.

Based on the data collected in Study 1, the probability of answering a comprehension question correctly given the lowest rating of perceived comprehension (the baseline) was estimated to be approximately 0.81. It was considered that central questions would be more likely to be answered correctly than peripheral questions (van den Broek & Helder, 2017; Yeari et al., 2017). As the centrality of the information targeted in comprehension questions was not manipulated in Study 1, a different value for the baseline probability was selected. As the expected difference in difficulty between semantically central and peripheral comprehension questions was unknown, a difference in probability of 0.1 was selected for convenience, centred around the baseline probability of Study 1. The baseline probability of correctly answering a central comprehension question was, therefore, set to 0.86 and the baseline probability of answering a peripheral question was set to 0.76. Solving the logit function yields β_0 = 1.82 and β_2 (capturing the change in probability from baseline, given a peripheral question) = -0.66, approximately.

Additional sources of variance on the baseline probability of answering a semantically central or peripheral comprehension question correctly were added to reflect the multilevel structure of the design. Magnitudes of variability similar to those observed in Study 1 were selected:

$$u_{0i} \sim \mathrm{N}(0, 1)\,,$$

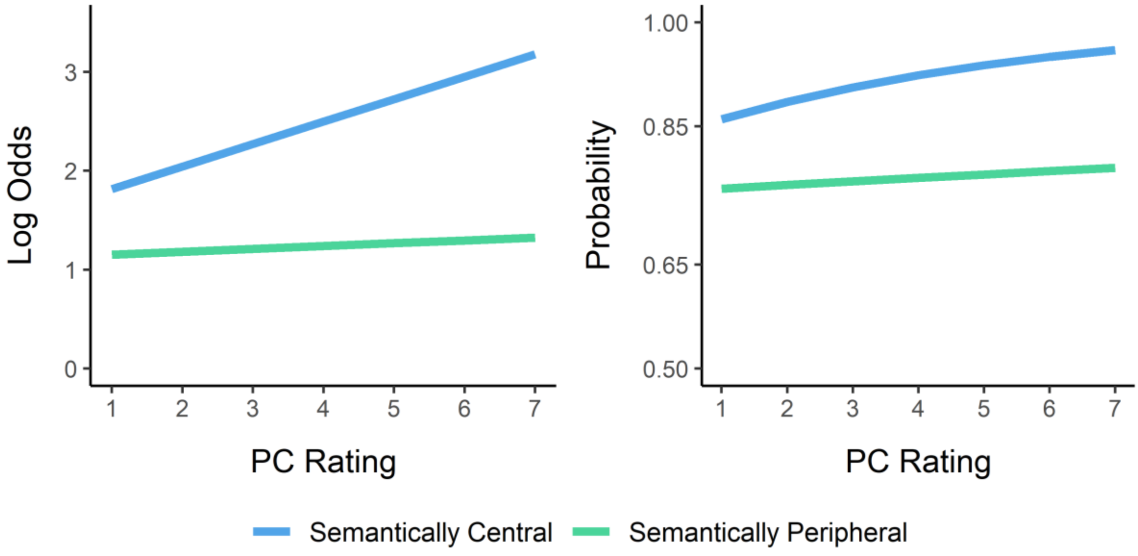
$$v_j \sim \mathrm{N}(0, 0.5)\,,$$

$$w_k \sim \mathrm{N}(0, 1)\,.$$

A conservative magnitude was selected to express the difference in the effect of perceived comprehension for semantically central and peripheral comprehension questions. A difference of +0.1 in the overall change in probability of a correct response was chosen between semantically central and peripheral information, given the minimum and maximum metacomprehension judgements. Magnitudes for β_1 and β_3 were calculated by centering the +0.1 difference around the estimated increase in the probability of a correct response observed in Study 1 (+0.08), from the minimum to maximum metacomprehension judgement. Therefore, the corresponding change in probability, given semantically central and peripheral comprehension questions, from the minimum to the maximum metacomprehension judgement, was selected as +0.13 and +0.03, respectively. Given the expanded seven-point rating scale and assuming a linear effect of metacomprehension judgements, solving the logit function yields β_1 = 0.23 and β_3 = -0.20, approximately. To illustrate this, Figure 4.1 shows the expectation of the log odds for the effect of perceived comprehension, given semantically central and peripheral questions.

Figure 4.1

Expectation of the Log Odds (Left) and Probability (Right) of a Correct Response in Study 2 Simulation



Note. PC = perceived comprehension.

Participant-level variability in the population-level effects were selected based on the variability observed in Study 1:

$$u_{1i} \sim \mathrm{N}(0, 0.1)\,,$$

$$u_{2i} \sim \mathrm{N}(0, 0.1)\,,$$

$$u_{3i} \sim \mathrm{N}(0, 0.1)\,,$$

As all deviates were drawn from univariate normal distributions, the covariance between the sources of variance were not constrained. This approach was taken due to the considerable uncertainty in this parameter observed in Study 1 and lack of evidence to indicate what direction or magnitude would be appropriate for these parameters.

For $Judgement_{ij}$, a vector of observations for each individual *i*, of length *J*, was simulated as the sum of 6 Bernoulli trials per text, corresponding to the perceived comprehension judgements for each individual across the set of texts. To produce a similar pattern of ratings as those in Study 1, the expected probability of each Bernoulli trial was 0.8. Variability from the probability of success on each trial was simulated per participant i, distributed $\mathrm{N}(0, 1)$, and per text j, distributed $\mathrm{N}(0, 0.5)$. This generated seven possible values for $Judgement_{ij}$ = (0,1,2,3,4,5,6), corresponding to each of the seven rating responses available for metacomprehension judgements.

Observations of $Centrality_k$ were binary, such that $Centrality_k$ = 0 corresponded to a comprehension question assessing semantically central information, while $Centrality_k =$ 1 corresponded to a comprehension question assessing semantically peripheral information. The number of central and peripheral comprehension questions were balanced within texts, with each text having *K*/2 of each type of question.

Simulation Procedure. The simulation was conducted using the HEC facility at Lancaster University using R (R Core Team, 2019). The simulation was conducted under varying numbers of participants, texts and questions per text. Five values were considered for the total number of participants *I* = (75, 100, 125, 150, 175), two values were considered for the total number of texts *J* = (10, 13), and two values were considered for the total number of questions per text K = (4, 6). A dataset of observations for each combination of participant, text and question sample sizes was simulated 1000 times.

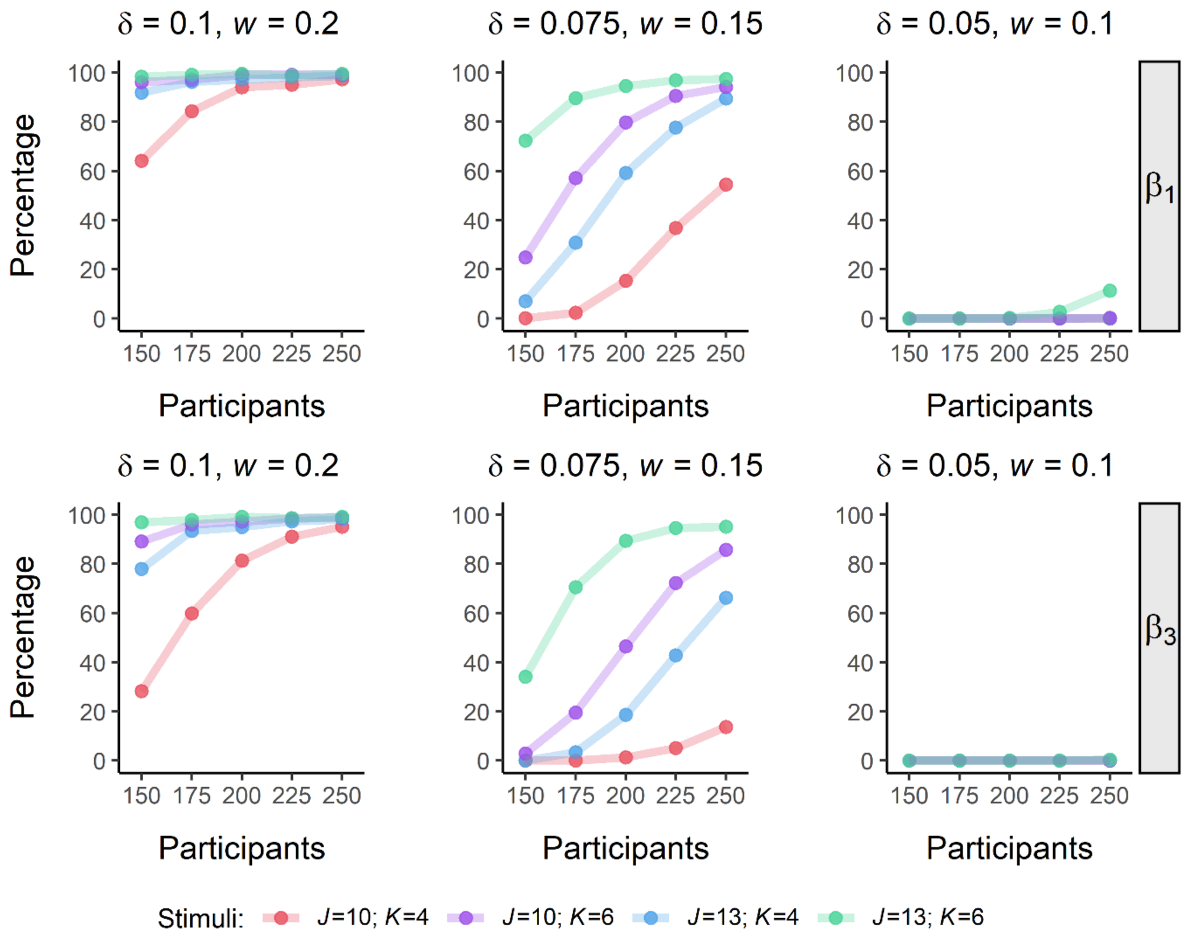
Model Fitting. The multilevel logistic model defined in (9) was fitted to each simulated dataset to estimate the effects of perceived comprehension for semantically central and for peripheral information using a Bayesian estimation framework, via the brms package (Bürkner, 2017, 2019) and Stan (Carpenter et al., 2017). Weakly informative priors were specified for the model parameters. For the population-level effects (β_0, β_1, β_2, β_3), normal distributions with mean 0 and standard deviation 10 were specified. Half-student-t distributions with the values of the degrees of freedom, location and scale parameters set to 3, 0 and 10 for the group-level effects ($u_{0i}, u_{1i}, u_{2i}, u_{3i}, v_j, w_k$). An LJK correlation distribution with a shape parameter of 1 was used as the prior on the covariance between participant-level variance in the intercept and the effect of perceived comprehension. From each resulting model fit, estimates for the mean of the posterior distribution and the 90% credible intervals for the effects of interest were obtained.

Simulation Results. The capacity to estimate the effects of interest, for a given number of participants and texts, was operationalised in terms of the probability of estimating β_1 and β_3 with accuracy and precision. Three pairs of values for δ and w were selected, representing **lower** ($\delta = 0.1, w = 0.2$), **middle** ($\delta = 0.075, w = 0.15$) and **higher** ($\delta = 0.05, w = 0.1$) levels of accuracy and precision in estimation. For the **lower** level of accuracy and precision, successful estimation of the effects excludes point estimates of less than half of the true effect magnitude and, separately, excludes values of more than double the true effect magnitude from the 90% credible intervals. For the **higher** level of accuracy and precision, successful estimation of the effects excludes values of less than half and more than double the true effect magnitude from the 90% credible intervals. The **middle** level of accuracy and precision represents a mid-point between these two estimation outcomes.

Across each of the selected sets of values for δ and w, successful estimation of the effect produces a 90% credible interval which does not overlap with zero. The probability of successfully achieving accuracy and precision in estimation, across varying δ and w, for each total number of participants, texts and questions per text simulated, is shown in Figure 4.2.

Figure 4.2

Probability of Achieving Varying Levels of Accuracy and Precision in Estimation in Study 2 Simulation



Note. The lines show the percentage of simulations achieving the specified target level of accuracy (δ) and precision (w) under varying numbers of participants (horizontal axis). From left to right, the level of accuracy and precision in estimation increases. The coloured lines correspond to various stimuli sample sizes, with varying numbers of total texts J and questions per text K. The upper and lower rows of the plot correspond to β_1 and β_3, the population-level effect of perceived comprehension for semantically central information and the population-level change in the effect of perceived comprehension given semantically peripheral information, respectively.

From left to right in Figure 4.2, decreasing values of δ and w correspond to higher demands on accuracy and precision in estimation for β_1 (top row) and β_3 (bottom row). The simulation indicated that, for the **lower** and **middle** targets for accuracy and precision, increasing participant and stimulus sample sizes considerably increased the probability of achieving the goal of estimation for both β_1 and β_3. In contrast, the larger considered sample sizes were not sufficient to substantially influence the probability of achieving the **higher** target level of accuracy and precision (right upper and lower plots in Figure 4.2). Across all sample sizes simulated, the probability of achieving the goal of estimation was consistently lower for β_3 than β_1.

Using the same sample sizes for texts and questions per text as Study 1 (J = 10, K = 4; red lines in Figure 4.2), 200 participants would provide a 94% and 81% probability of achieving the **lower** level of accuracy and precision for β_1 and β_3, respectively. In contrast, given the **middle** level of accuracy and precision, none of the participant sample sizes considered would provide a high probability of achieving the estimation goal. This indicates that a greater volume of data is required to estimate with accuracy and precision in Study 2.

Either increasing the number of questions per text to six (purple lines, Figure 4.2) or simultaneously increasing both the number of texts and questions per text (green lines, Figure 4.2) provides a higher probability of observing the same level of accuracy and precision as targeted in Study 1. For β_3, given six questions per text, with 10 texts in total, 250 participants would provide an 86% probability of observing the **middle** target level of accuracy and precision. Given an additional three texts, however, 50 fewer participants provide an 89% chance of achieving the goal of estimation.

Based on the simulation, a total of 13 texts with six questions per text was selected for the design of Study 2. For the **middle** level of accuracy and precision, a total of 200 participants would yield a more than 80% probability of observing accurate and precise estimates. However, given the desire to jointly achieve the goal of estimation for both β_1 and β_3, in addition to balancing concerns including available resources, a larger participant sample size was selected. A participant sample size of 225 was selected to provide an approximately 92% probability of simultaneously estimating both β_1 and β_3 with the **middle** level of accuracy and precision in estimation.

4.2.2 Pilot

Given the similarity with the design of Study 1, a 'within data collection' pilot was conducted to evaluate the proposed study design. This consisted of pausing data collection after the first 25 participants had completed the metacomprehension task, checking responses to detect any major problems with the experimental apparatus and confirming that the duration of the experimental session was as anticipated. The first 25 participants were aged 18 to 73 (*M* = 37.88, *SD* = 16.48), with 19 identifying as female and six as male. No formal analyses were run on this subset of the sample. As no concerns were identified, data collection was resumed until the target number of participants had been reached. Since the recruitment process, materials and procedure were identical in the pilot and Study 2, these will be discussed in the context of the main data collection below.

4.2.3 Participants

A sample of 225 participants was recruited using the online platform Prolific. Participation was limited to UK nationals that had not taken part in any tasks in Study 1. Recruitment was conducted over several days to capture a greater spread of participants on the Prolific system. Consistent with Study 1 (see Chapter 3; section 3.2.3), exclusion criteria specified that participants would be excluded and replaced if they demonstrated limited evidence of engagement with the task. Reading times were used as a proxy measure of engagement, with insufficient engagement defined as reading times less than the minimum duration given a reading rate of 600 wpm. The impact of participant exclusion on the results are explored in the sensitivity analysis (section 4.3.3).

Given the length of the testing sessions, participation in the experimental tasks was split into two parts. The first session consisted of the metacomprehension task (reading, judging and answering comprehension questions from health-related texts) and the second session consisted of the individual difference tasks (measuring reading ability, background knowledge and standards of coherence). Participants were recruited to complete the first session and were invited to complete the second session approximately one week after the first session. Only participants who fully completed the first session were invited to complete the second session. All participants, including those who met the exclusion criteria, were invited to complete the second session. The second session remained available to complete for approximately two weeks. All participants received £3.00 for completing each section of the study.

Following the sampling procedure outlined, participants were recruited to complete the first session until 225 submissions which met the acceptance criteria were obtained (page submission times on all 13 health-related texts corresponding to a reading rate greater than 600 wpm). In total, 300 participants completed the first session. Responses from 75 participants who met the exclusion criteria were removed and replaced. Overall, a low rate of attrition between sessions was observed: 95% of participants returned to complete the second session.

Of the 300 participants recruited to the first session, the majority were young and female. The average age was 36.22 (*SD* = 13.83), with 126 participants reporting as male, 172 as female, and two as non-binary. The sample characteristics of participants meeting the acceptance criteria were similar. The average age of the 225 accepted participants was 37.80 (*SD* = 14.22), with 99 participants reporting as male, 125 as female and one as non-binary. This comparability was also true of the participants who returned. The average age of the 284 participants who returned to complete the second session was 38.06 (*SD* = 14.35), with 94 participants reporting as male, 119 as female, and one as non-binary.

4.2.4 Materials

Health Texts.

Sampling Procedure. To collect a sample of texts likely to be more diverse in difficulty than those used in Study 1, online sources of health information beyond NHS webpages were targeted. Text sampling was done in a two-step approach: 10 texts were initially selected and text features were evaluated, followed by the selection of a further three. An opportunity sample of ten outlets of online health information was collected via searching the web (health conditions examined in Study 1 were used as search terms). To avoid inaccurate health information, only outlets which declared the involvement of medical expertise in the production of the information were included. Outlets which were fully pay-walled were excluded. From each outlet, one text topic was selected.

To select text topics, 10 conditions were randomly identified, using a random letter and number generator, from the A-Z of rare diseases listed on rarediseases.org. Topics were selected by iteratively randomly selecting a topic from the list and then verifying its suitability. Suitability was assessed according to whether the topic was appropriate (not likely to be an embarrassing or distressing condition), featured on the target outlet website, and whether there was sufficient text to create a 200-300 word stimulus text.

After the selection of 10 text topics, stimuli were constructed from the online information (full description in the following subsection). Descriptive information for the constructed stimuli was then generated using the Coh-Metrix 3.0 online tool (Graesser, et al.,

2004) to evaluate text features of interest. This roughly indicated there was a cluster of seven texts which scored more highly on readability measures, compared to three lower readability texts. To provide more opportunities to observe lower ratings of perceived comprehension, an additional text was obtained from each of the three outlets from which the lower readability texts were constructed, selected using the sample strategy described above, producing the full set of 13 health-related texts.

Table 4.1

Descriptive Information for the Health Texts in Study 2

Topic	Words	Flesch Score	Fleisch-Kincaid
Porphyria	273	61.3	8.47
Scleroderma	257	55.8	8.51
Neutropenia	214	9.4	15.27
Seborrhoeic dermatitis	228	59.6	8.49
Erysipelas and Cellulitis	241	51.4	9.24
Cholestasis of pregnancy	268	60.5	7.48
Dengue Fever	286	51.1	9.36
Bilateral renal agenesis	243	60.7	8.35
Zollinger-Ellison syndrome	247	31.9	12.65
Myasthenia gravis	252	29.8	14.32
Coxiella burnetii infection	231	8.03	16.24
Sialadenitis	257	42.79	10.07
Brucellosis	285	38.38	12.96

Note. Flesch score refers to the Flesch Reading Ease score. Flesch-Kincaid refers to the Flesch-Kincaid Grade level. Calculated using the online tool Coh-Metrix 3.0 (Graesser, et al., 2004).

Stimulus Construction. For each health topic, sections of text from the outlet webpage were taken to produce 200-300 word texts. Links, images and formatting were removed, but heading and subheadings were preserved. Sections which were predominantly lists, or discussed aspects likely to cause distress, were omitted. The coherence of the resulting text was checked, by rereading, to ensure that flow and meaning was preserved. American English spellings were changed to British English. Descriptive information for the texts is presented in Table 4.1 and the full texts are provided in Appendix I.

Comprehension Questions. For each text, six comprehension questions were constructed: three targeting semantically central information and three targeting semantically peripheral information. To construct these questions, information which varied in semantic centrality was identified from the texts. Given the variable approaches in defining and quantifying centrality, an attempt was made to create a procedure to identify central and peripheral elements with greater objectivity and control. The approach incorporated ideas from multiple sources, concerning the nature and features of central and peripheral information (e.g. Helder et al., 2019; Kintsch, 1988; Thorndyke, 1977; van den Broek et al., 2013; van den Broek & Helder, 2017; van Dijk & Kintsch, 1983). In the development and application of the approach, however, it was recognised that i) the process is not an exhaustive measure of the ways in which an idea in a text may be more or less central, and ii) a degree of subjectivity was unavoidable (owing to the difficulty in operationalising some concepts, such as connections and elaborations).

For each text, information was divided into coarse idea units, expressing a single idea, which were then used to construct a semantic map of the connections between ideas in the text. The semantic maps were jointly constructed by two researchers, with disagreements resolved through discussion. From the semantic map, information which was more versus less-well connected, representing semantically central versus peripheral information, was selected (see Appendix J for details). As not all information was amenable to question construction, some idea units were not considered suitable for use. For example, some idea units contained only numerical information (e.g., “40,000 are affected”) which presented a challenge in avoiding testing only surface level processing. Following this procedure, three semantically central and three semantically peripheral idea units were identified for each text. The full procedure is detailed in Appendix J.

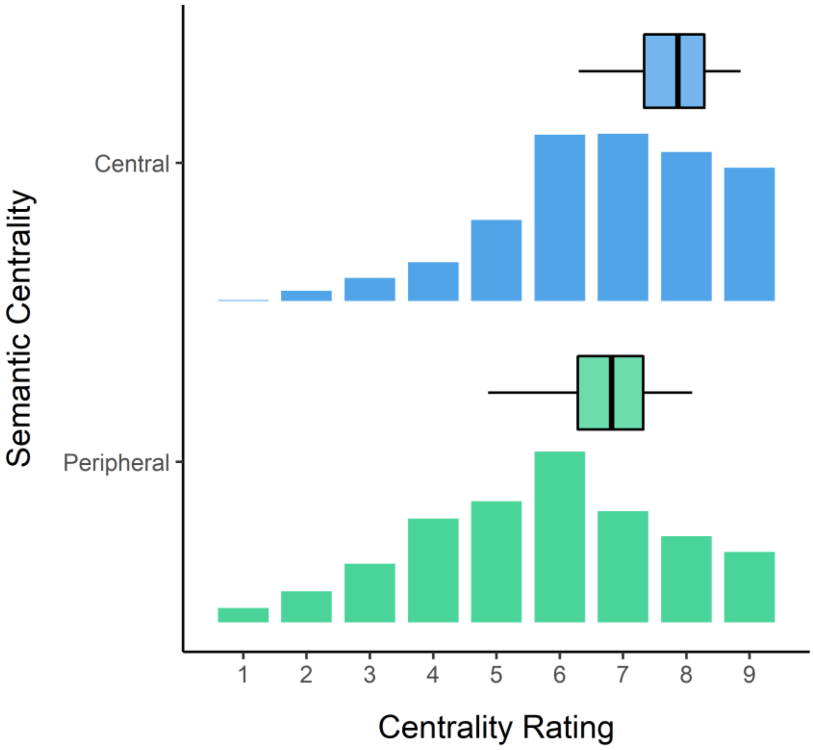
Evaluation of Idea Unit Centrality. After identifying 78 idea units, a small-scale study was conducted to collect ratings of centrality. This information was obtained to examine the relationship between the selected idea units and readers' perceptions of centrality. In addition, these ratings were collected to provide an alternative measure of centrality for use in the sensitivity analysis. To collect centrality ratings, participants were recruited via Prolific. To try to ensure centrality ratings were informed by adequate comprehension of the text, the sample pool was restricted to individuals who had completed higher education. In addition, participants produced a brief written summary of each text which was evaluated to detect evidence of misunderstanding (the ratings of two participants were excluded on this basis). Participants were recruited until there were at least 20 centrality ratings for each of the 78 idea units (ratings from participants who provided partial data were also included). The sample of participants providing centrality ratings consisted of 27 individuals, 19 reporting as female and eight as male, with an average age of 37.74 (*SD* = 14.01). Participants received £2.50 for taking part.

In the centrality rating task, each text was presented on screen, followed by each of the six idea units per text, formed into coherent sentences. Participants were instructed to read the texts for the purpose of obtaining a general understanding of the condition. The rating scale appeared as an unmarked line with the labels 'Less important' on the left and 'More important' on the right of the scale. Ratings were elicited with the prompt 'Please rate the importance of the statements in understanding the text'. Previous research has utilised a variety of scale sizes to elicit centrality ratings, including four (Johnston & Afflerbach, 1982), five (Yeari et al., 2015; Yeari et al 2017), seven (Albrecht & O'Brien, 1991; Mo et al., 2007) or eight option response scales (Miller et al., 2013). As there is no clear standard, a 9-point scale was chosen to provide reasonable reliability and response discrimination (Dawes, 2008; Leung, 2011; Preston & Colman, 2000). Additional guidance was provided at the start of the experiment to clarify the rating task. These instructions stated: 'You will judge how important each statement is in understanding the health condition. You might find it helpful to read the text and try to summarise it in your head before rating. To make a rating, you could think about: if you had to explain the condition to someone else, how important it would be to include the statement in your summary; whether the text would make sense if the information in the statement was not in the text; how well connected the statement is to other parts of the text'.

The relationship between the idea units and readers' perceptions of centrality was evaluated through data visualisation. The distributions of rated semantic centrality, for idea units identified as semantically central or peripheral, is shown in Figure 4.3. Idea units identified as central were typically rated as more semantically central, on average, than those identified as peripheral. However, substantial within-classification variability was observed, with the ratings of some idea units incongruous with their identification. As the impact of differences in operationalising centrality was of interest, no changes to the 78 idea units were made based on these responses. The impact of this alternative measure of semantic centrality is explored in the sensitivity analysis (section 4.3.3).

Figure 4.3

Ratings of Semantic Centrality for Study 2 Idea Units



Note: The boxplots show the median ratings of centrality, aggregated across participants, for each of 34 idea units classified as central (blue) and each of the 34 idea units classified as peripheral (green). The histograms show the ratings of centrality aggregated across all participants and idea units, by centrality identification.

Multiple-choice questions were developed to assess understanding of information of the 78 idea units identified. It was considered that assessing understanding of the idea unit itself, rather than of an inference generated on the basis of it, was of primary relevance. Idea units captured individual pieces of information which, while they could be used to form inferences, were not derived from inferences themselves. As such, an alternative approach to constructing inference-based questions was taken: questions were designed to test comprehension of the information corresponding to the idea unit which was explicit in the text, while avoiding surface overlap. Identifying the correct answer required matching a semantically accurate, but lexically dissimilar, response option with the information provided in the text. Three distractors were created for each question item. Distractors primarily consisted of response options with high lexical overlap with the text, likely misconceptions, or semantic near-misses. Constructed questions and response options were separately reviewed by three experimenters and altered until all experimenters were satisfied. The final set of questions with response options is provided in Appendix K.

Perceived Comprehension. Consistent with Study 1, judgements elicited using the prompt 'How much of the text do you feel you understand?' provided the measure of perceived comprehension in the main analysis. To allow for comparisons of the estimated predictive relationship between metacomprehension judgements and assessed comprehension between various judgement prompts, the four additional prompts used in Study 1 were also presented (see Chapter 3, section 3.2.4).

Three alterations were made to the rating scale used in Study 1 to capture judgements. First, responses were captured using a 7-point scale. Second, given the expanded scale, only labels on the ends of the response scale line were shown, rather than at all points within the scale. The same labels for the end points of the scales used in Study 1 were displayed for each prompt (see Chapter 3, section 3.2.4). Thirdly, given the different presentation of labels, the response scale was presented as an unmarked line with a moveable slider, initially positioned at the mid-point on the scale, to be dragged left or right to indicate judgement (the slider ‘snapped’ to the value closest to one of the seven ratings).

QRI. There were no changes to the administration of this task between Study 1 and Study 2. Comprehension performance on the two ‘life cycle of stars’ passages, taken from the QRI, provided a measure of reading ability. Participants provided open-ended responses to the 20 questions which were marked using the altered version of the QRI marking rubric developed in Study 1. Two markers graded responses as correct (1) or incorrect (0). Intermarker agreement was generally high (percentage agreement per question ranged from 90% to 100%, *M* = 96%). Each disagreement was resolved through discussion.

HLVA. The 16-item adaptation of the HLVA used in Study 1 provided a measure of health-related background knowledge. Participant responses to multiple-choice questions (Appendix D) with full scoring applied. There were no changes to the presentation of this task between Study 1 and Study 2.

Reader-Based Standards of Coherence Factor 3 (RBSOCF3). Given the potential role of standards of coherence in shaping the macrostructure constructed following reading, to measure individuals' dispositional desire to understand what they are reading, participants completed the third subscale of Calloway's (2019) measure of reader-based standards of coherence. This subscale attempts to capture an individual's motivation to understand text and use of regulation strategies, which are argued to most closely tap their standards of coherence (Calloway, 2019; van den Broek et al., 1995, 2001). This measure consists of nine items, with responses collected using a 7-point Likert scale labelled at the extremes as 'strong disagreement' and 'strong agreement'. Responses to each item are summed to provide a score for the reader's desire to understand.

4.2.5 Procedure

Participation in the study commenced when participants responded to the online invitation on Prolific.co. All tasks were computer-based and presented via Qualtrics. To reduce the potential for fatigue induced by long testing sessions, participants completed the metacomprehension task in one session and the individual differences measures (QRI, HLVA, RBSOCF3) in a second session approximately one week later. Participants were informed of the second session, provided consent and reported their age and gender prior to completing the tasks.

In the first session, participants read each health text in the metacomprehension task and made judgements of comprehension immediately following the reading of each text. Instructions provided before reading and judging the texts were the same as in Study 1. Participants were invited to take a break if needed after reading and judging all texts. Participants were then presented with the comprehension questions for each text, with the text present, in the same order that the texts were presented for reading. The presentation order for each text, comprehension question and response option were randomised. After completing the comprehension questions, participants were reminded to return to Prolific.co for the second session. In the second session, participants completed the QRI, HLVA and RBSOCF3. The tasks were presented in randomised order.

4.3 Results

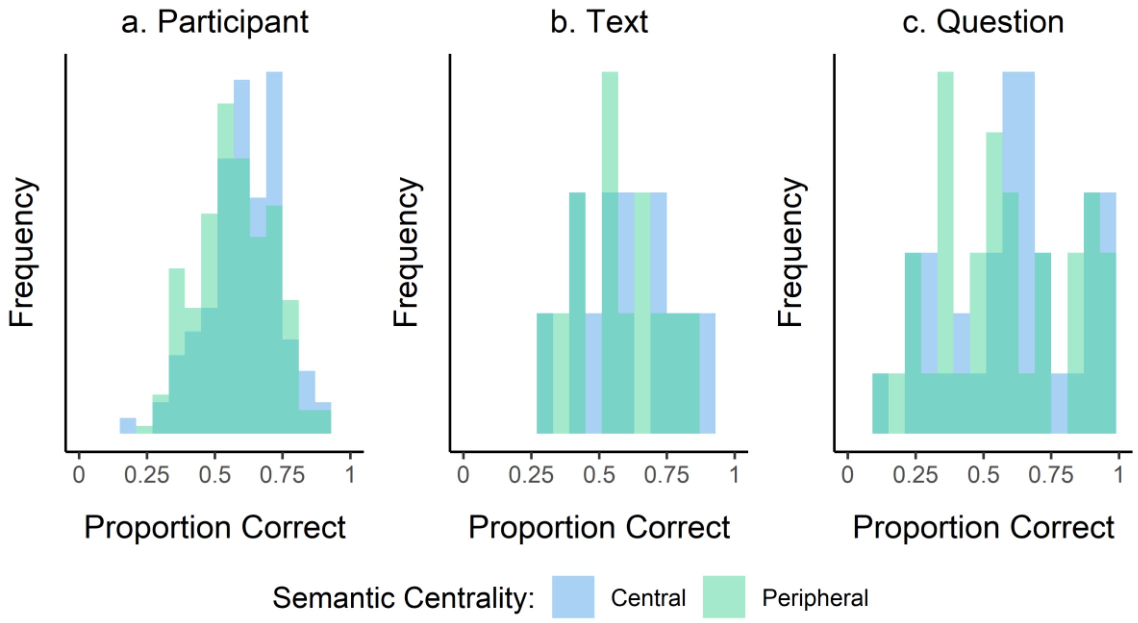
4.3.1 Preliminary Data Inspection

Initial inspection of the responses which met the acceptance criteria indicated that performance on comprehension questions was lower than Study 1. Across all participants, texts and questions, the proportion of comprehension questions answered correctly was 0.61 and 0.57 for questions targeting semantically central and semantically peripheral information, respectively. This corresponds to participants correctly answering between three to four of the total six questions per text, on average. Variability in this was observed at the level of the participant, text and question.

Figure 4.4 shows the distributions of the proportion of correct responses by participant (4.4a.), text (4.4b.), and question (4.4c.), coloured according to semantically centrality. Variability in proportion correct was comparably greater between questions.

Figure 4.4

Proportion of Semantically Central and Peripheral Comprehension Questions Correctly Answered in the Metacomprehension Task in Study 2

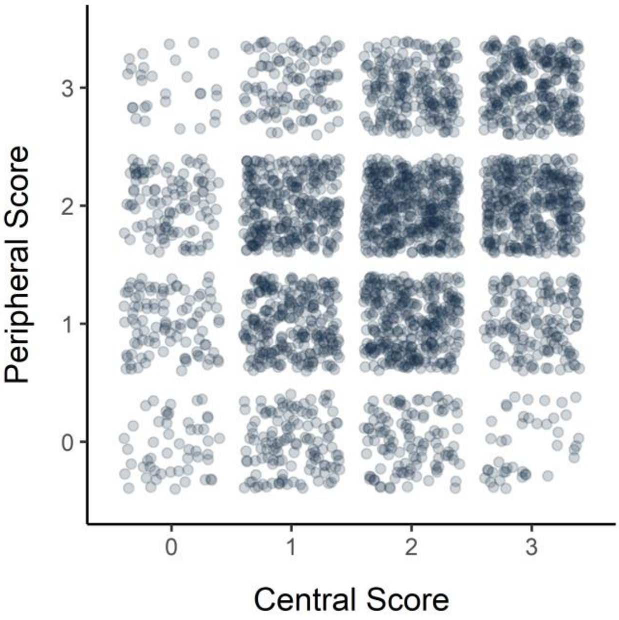


Note. Participant proportion correct (*I* = 225), text proportion correct (*J* = 13), and question proportion correct (*K* = 78). Proportion correct are coloured by semantic centrality: overall response accuracy on semantically central and semantically peripheral comprehension questions is shown in blue and green, respectively.

In addition, considering the potential impact of a high association between performance on semantically central and semantically peripheral comprehension questions on estimating the predictiveness of metacomprehension judgements, the relationship between performance on central and peripheral comprehension questions was examined. The number of comprehension questions a participant answered correctly was summed by text and according to semantic centrality. The relationship between summed performance on questions assessing comprehension of semantically central and peripheral information is illustrated in Figure 4.5, indicating a positive association between the text-level sum scores across participants.

Figure 4.5

Scatterplot of Central and Peripheral Comprehension Performance in Study 2

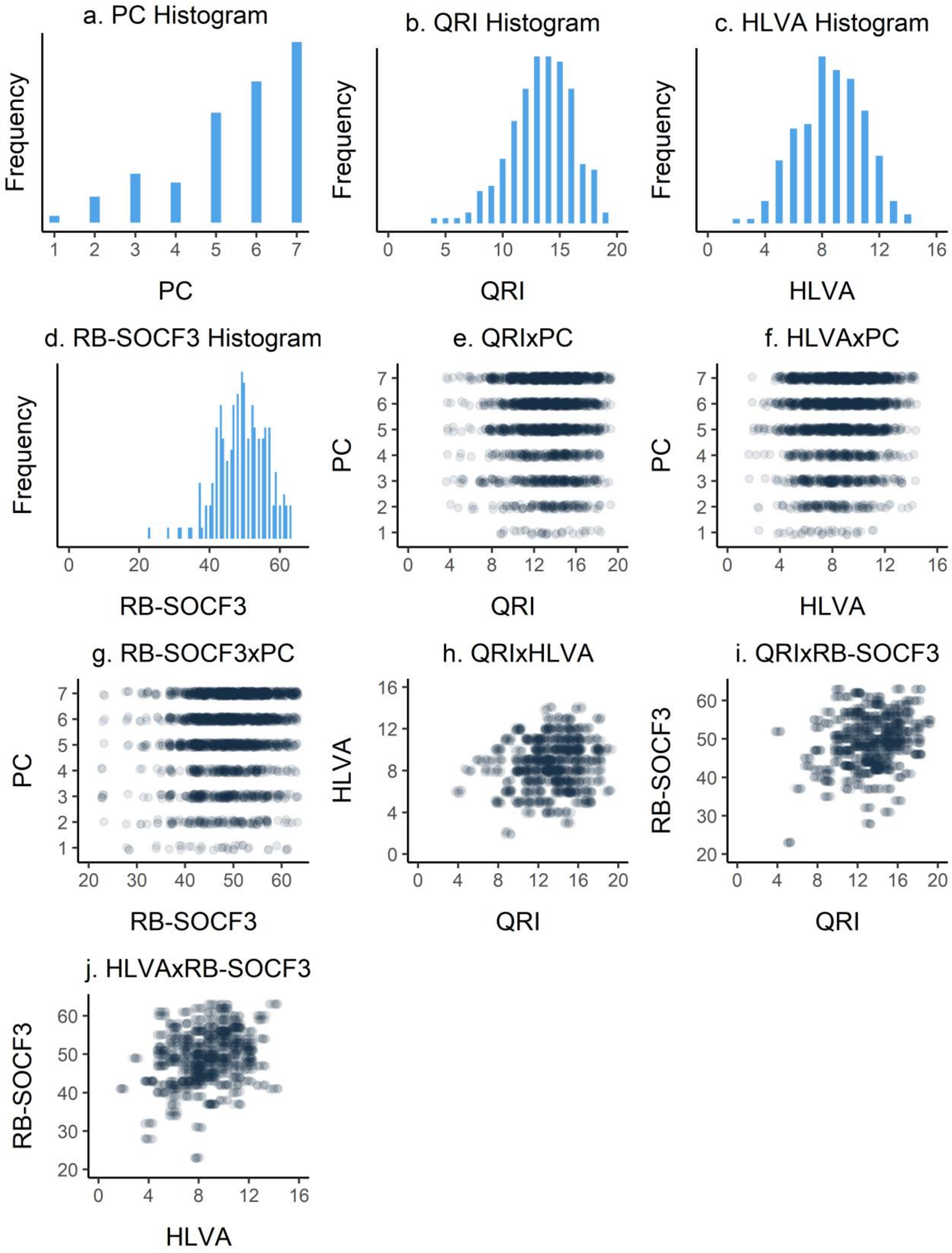


Note. Points are plotted with jitter applied, adding small, randomised perturbations to the observations to increase the readability of the plot. Central and peripheral score refer to the number of comprehension questions correctly answered by an individual per text.

The distributions of variables to be used as predictors in the main analysis (perceived comprehension judgements, QRI scores, HLVA scores, and RBSOCF3 scores) and the corresponding bivariate scatterplots are shown in Figure 4.6. Responses to the perceived comprehension prompt selected for use in the main analysis: ‘How much of the text do you feel you understand?’ were negatively skewed. As can be seen in Figure 4.6a, participants predominantly selected the three highest ratings when providing metacomprehension judgements. Differences in judgement distributions were observed between texts (see Figure L.1 in Appendix L). All rating values featured in the response data. The scatterplots of perceived comprehension and QRI scores (4.6e), HLVA scores (4.6f), and RBSOCF3 scores (4.6g) indicated no clear bivariate association between these measures.

Figure 4.6

Histograms and Scatterplots of the predictor variables considered Study 2



Note. Points in scatterplots e.-j. are plotted with jitter applied, adding small, randomised perturbations to the observations to increase readability. PC = perceived comprehension, QRI = Qualitative Reading Inventory, HLVA = Health Literacy Vocabulary Assessment, RBSOCF3 = Reader-Based Standards of Coherence Factor 3.

Across participants, performance on both the HLVA (*M* = 8.62, *SD* = 2.29), QRI (*M* = 13.37, *SD* = 2.73) and RBSOCF3 (*M* = 49.24, *SD* = 6.89), shown in Figure 4.6b, 4.6c and 4.6d, respectively, were all broadly normally distributed with some negative skew. These distributions indicate that the sample captured a range of different levels of reading ability, health-related background knowledge, and dispositional standards of coherence. The scatterplots for these variables, shown in Figure 4.6h, 4.6i, 4.6j, suggest weakly positive associations amongst these three measures.

4.3.2 Planned Analysis

To explore whether judgements of comprehension are better predictors of the comprehension of semantically central or peripheral information on health-related texts, and whether individual differences in reading ability, background knowledge and standards of coherence might moderate such a predictive relationship, Bayesian multilevel logistic regression models were fitted in R (R Core Team, 2019) using the brms package (Bürkner, 2017, 2019) and Stan (Carpenter et al., 2017). The results of the analyses corresponding to each research question are discussed below. Following this, a sensitivity analysis is reported to explore how various analytical choices may influence the findings.

RQ3. To address RQ3, the model defined in (9), repeated below for convenience, was fitted to the data using a Bayesian estimation framework:

$$\log\left(\frac{P_{ijk}}{1-P_{ijk}}\right) = \beta_0 + (\beta_1 + u_{1i})Judgement_{ij} + (\beta_2 + u_{2i})Centrality_k \qquad (9)$$

$$+(\beta_3 + u_{3i})Judgement_{ij}Centrality_k + u_{0i} + v_j + w_k,$$

Observations of assessed comprehension, from individual *i* on text *j* and question *k*, were used as the outcome measure Y_{ijk}. Judgements of perceived comprehension, standardised prior to model fitting, were entered as predictors of the log odds of correctly responding to the comprehension question. Group-level variance in the intercept was estimated for participants u_{0i}, texts v_j, and questions w_k. Participant-level variability in the three population-level effects was included: u_{1i}, u_{2i}, u_{3i}. The same weakly informative priors described in the design analysis (section 4.2.1) were used for the population-level effects, group-level effects, and covariance parameters.

The model was estimated using six chains with 8000 iterations each, half of which were discarded as burn-in. To pre-empt divergent transitions, 'adapt_delta' was set to 0.99. The model appeared to converge well under this specification, however, a small number of parameters relating to participant-level variability in the effect of question type were estimated with comparably lower effective sample sizes than other model parameters. Given that low numbers of effective samples may produce unreliable estimates for posterior means and interval values, the estimate for this variance parameter should be considered cautiously. The full results of this model are presented in Table 4.2.

The mean of the posterior distribution for the effect of perceived comprehension judgements was estimated to be positive. Given the binary coding of the semantic centrality indicator variable, this indicates that the log odds of correctly answering a comprehension question targeting semantically central information increases by 0.08 per unit increase in perceived comprehension, on average. For semantically central information, both the 95% credible interval (see Table 4.2) and 80% HDI (80% HDI = [0.03, 0.13]) for the effect of perceived comprehension was located over positive values. The change in the log odds (interaction), given a comprehension question assessing semantically peripheral information, was estimated to be an additional increase on this effect of 0.02, on average, per unit increase in perceived comprehension. However, the direction of the change in the effect, for semantically peripheral information, is less clear. For this effect, both the 95% credible interval and the 80% HDI span zero (80% HDI = [-0.04, 0.09]).

Table 4.2

Bayesian Multilevel Logistic Model of Response Accuracy in Study 2

Parameter	Estimate[a]	Error[b]	95% CI[c]	Eff Sample
Population-level Effects				
Intercept	0.66	0.29	[0.08, 1.24]	6009
Perceived comprehension	0.08	0.04	[0.01, 0.16]	20115
Semantic centrality	-0.23	0.28	[-0.77, 0.33]	4436
Perceived comprehension x semantic centrality	0.02	0.05	[-0.07, 0.12]	28982
Group-Level Variance				
Participant (intercept)	0.65	0.04	[0.57, 0.73]	8438
Participant (perceived comprehension)	0.12	0.05	[0.02, 0.21]	2671
Participant (semantic centrality)	0.15	0.08	[0.01, 0.30]	1838
Participant (perceived comprehension x semantic centrality)	0.06	0.05	[0.00, 0.18]	5460
Text (intercept)	0.72	0.27	[0.26, 1.31]	3270
Question (intercept)	1.21	0.11	[1.01, 1.45]	5531
Covariance of intercept and slope variance				
Intercept, perceived comprehension	0.04	0.25	[-0.48, 0.53]	13059
Intercept, semantic centrality	-0.02	0.30	[-0.56, 0.63]	12801
Intercept, perceived comprehension x semantic centrality	0.02	0.39	[-0.75, 0.76]	30050
perceived comprehension, semantic centrality	-0.49	0.37	[-0.94, 0.48]	3061
perceived comprehension, perceived comprehension x semantic centrality	-0.12	0.45	[-0.86, 0.77]	18226
semantic centrality, perceived comprehension x semantic centrality	-0.01	0.45	[-0.82, 0.81]	13481

Note: Population-level effect estimates are presented in logits. Rhat values for all parameters = 1.00. CI = credible interval. Eff Sample = number of effective samples, obtained using the bayestestR package (Makowski et al., 2019).

[a]Estimate refers to the mean of the marginal posterior distribution of the parameter. [b]Error refers to the

standard deviation of the marginal posterior distribution of the parameter. [c]Credible intervals represent the upper and lower values within which 95% of the estimated parameter values in the posterior distribution are contained.

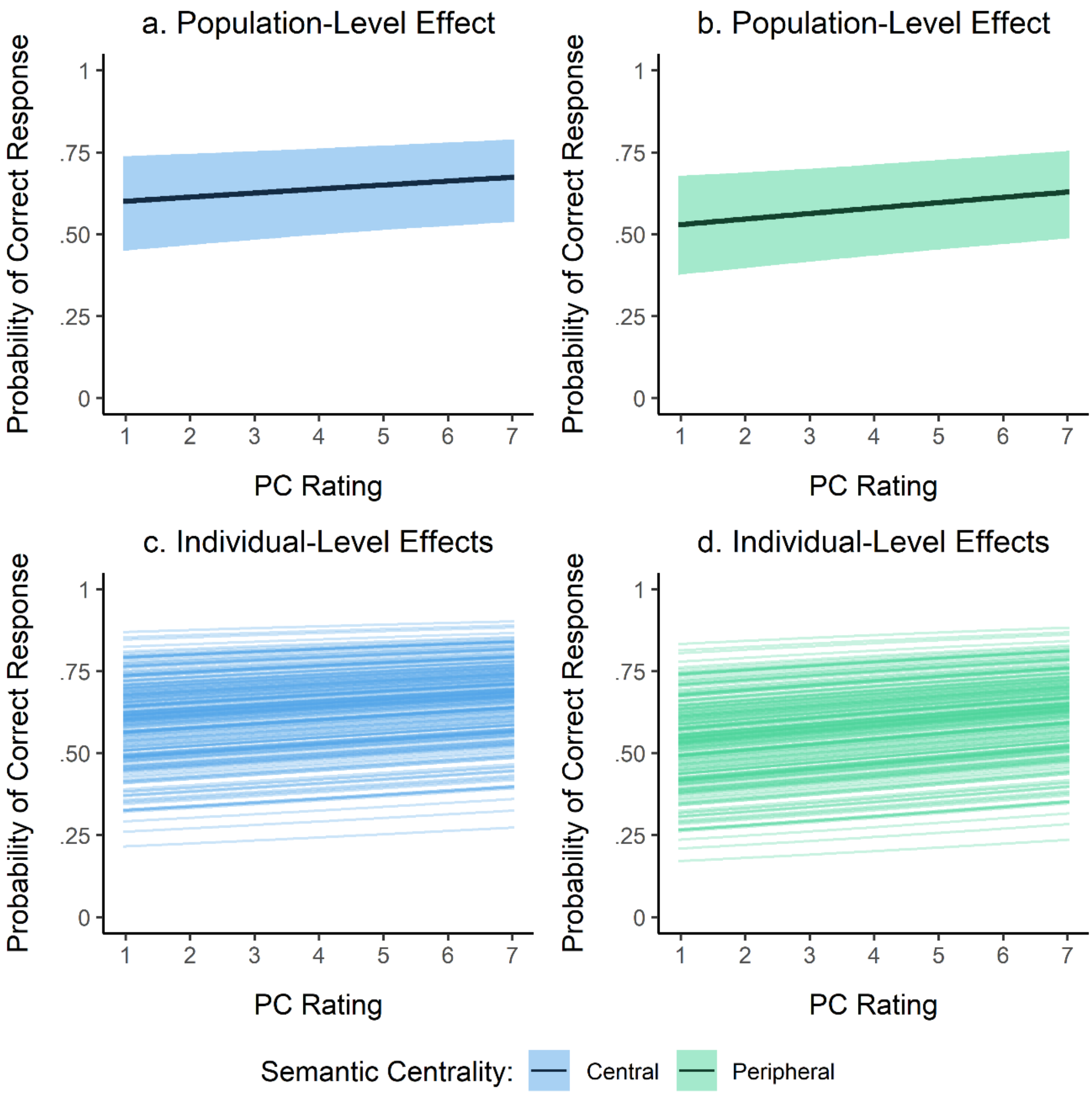
Based on model-fitted predictions of the marginal effects, the lowest and highest perceived comprehension judgements would correspond to an expected probability of answering a semantically central comprehension question correctly of 0.60 (*SE* = 0.07) and 0.67 (*SE* = 0.06), respectively. In contrast, for semantically peripheral questions, these probabilities would be 0.53 (*SE* = 0.08) and 0.63 (*SE* = 0.07), respectively. The estimated uncertainty in these effects can be seen in Figure 4.7, showing the marginal model-fitted predictions of correctly responding to questions assessing semantically central (4.7a) and peripheral information (4.7b) across the perceived comprehension judgement scale.

Participant-related variability in the predictive relationship between perceived comprehension judgements and performance on semantically central and peripheral comprehension questions was estimated to be limited. Participant-level model-fitted estimates of the effect of perceived comprehension shown in Figure 4.7c and 4.7d, for semantically central and peripheral information respectively, demonstrate that the effect remained small and positive across individuals.

The posterior distributions for covariance parameters typically featured close-to-zero means and large standard deviations. A sizable negative posterior mean was estimated for the correlation between participant-level deviances in the effect of perceived comprehension and the effect of semantic centrality. However, large uncertainty was associated with this parameter (see Table 4.2).

Figure 4.7

Estimated Effect of Perceived Comprehension, at the Population- and Individual-level, by Semantic Centrality, in Study 2



Note. Marginal model-fitted predictions at the population-level (a. and b.) and conditional model-fitted predictions at the individual-level (c. and d.) are plotted, coloured according to whether the correct response corresponds to semantically central (blue) or peripheral (green) information. Panels a. and b. show the expected probability of a correct response for the 'average' participant, responding to the 'average' question concerning an 'average' text, with the 95% credible interval shaded. Panels c. and d. show the expected probability of a correct response for each participant, given an 'average' question concerning an 'average' text. PC = perceived comprehension.

Differing magnitudes of group-level variability in the overall probability of answering a comprehension question correctly were estimated across participants, texts, and questions, reflecting observed variability in response accuracy (Figure 4.4). Text-level differences contributed to variability in response accuracy to a slightly greater extent than individual-level differences. Differences between questions, however, remained the largest source of variability in response accuracy (see Table 4.2).

RQ4. To address RQ4, the regression model defined in (9) was extended to estimate the effects of QRI score β_4, HLVA score β_5, RBSOCF3 score β_6 and the interactions between QRI score and perceived comprehension β_7, HLVA score and perceived comprehension β_8, and RBSOCF3 score and perceived comprehension β_9:

$$\begin{aligned}\log\left(\frac{P_{ijk}}{1-P_{ijk}}\right) = {} & \beta_0 + (\beta_1 + u_{1i})Judgement_{ij} \ + (\beta_2 + u_{2i})Centrality_k \\ & +(\beta_3 + u_{3i})Judgement_{ij}Centrality_k + \beta_4 QRI_i + \beta_5 HLVA_i \\ & +\beta_6 RBSOCF3_i + \beta_7 Judgement_{ij}QRI_i + \beta_8 Judgement_{ij}HLVA_i \\ & +\beta_9 Judgement_{ij}RBSOCF3_i + u_{0i} + v_j + w_k \ .\end{aligned} \tag{10}$$

Judgements of perceived comprehension, QRI scores, HLVA scores, and RBSOCF3 scores were standardised prior to model fitting. The same weakly informative priors described in the design analysis (section 4.2.1) were used for the population-level effects, group-level effects, and covariance parameters. For the six additional population-level effects ($\beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9$), normal distributions with mean 0 and standard deviation 10 were specified as priors.

The model was estimated using six chains with 8000 iterations each, half of which were discarded as burn-in. To pre-empt divergent transitions, ‘adapt_delta’ was set to 0.99. The model appeared to converge well under this specification. However, as with the fitting of model (9), a small number of parameters, relating to participant-level variances in the effect of question type, were estimated with comparatively lower effective sample sizes than other model parameters. The estimate for this variance parameter should, therefore, be considered with additional uncertainty. The full results of this model are presented in Table 4.3.

Table 4.3

Extended Bayesian Multilevel Logistic Model of Response Accuracy in Study 2

Parameter	Estimate[a]	Error[b]	95% CI[c]	Eff Sample
Population-level Effects				
Intercept	0.69	0.30	[0.10, 1.28]	8645
Perceived comprehension	0.07	0.04	[-0.01, 0.15]	27941
Semantic centrality	-0.23	0.28	[-0.78, 0.33]	5821
Perceived comprehension x semantic centrality	0.03	0.05	[-0.06, 0.13]	33677
QRI	0.16	0.04	[0.08, 0.24]	10558
HLVA	0.28	0.04	[0.20, 0.36]	11398
RBSOCF3	0.06	0.04	[-0.03, 0.14]	10218
Perceived comprehension x QRI	-0.04	0.02	[-0.09, 0.00]	31845
Perceived comprehension x HLVA	0.03	0.02	[-0.02, 0.08]	33489
Perceived comprehension x RBSOCF3	0.02	0.02	[-0.03, 0.07]	24887
Group-Level Variance				
Participant (intercept)	0.52	0.04	[0.45, 0.60]	11918
Participant (perceived comprehension)	0.12	0.05	[0.01, 0.21]	2842
Participant (semantic centrality)	0.16	0.08	[0.01, 0.31]	2189
Participant (perceived comprehension x semantic centrality)	0.07	0.05	[0.00, 0.20]	5413
Text (intercept)	0.74	0.27	[0.26, 1.36]	3676
Question (intercept)	1.21	0.12	[1.01, 1.47]	6388
Covariance of intercept and slope variance				
Intercept, perceived comprehension	-0.01	0.27	[-0.57, 0.51]	12918
Intercept, semantic centrality	0.06	0.30	[-0.49, 0.69]	13522
Intercept, perceived comprehension x semantic centrality	0.12	0.38	[-0.68, 0.80]	30762

perceived comprehension, semantic centrality	-0.46	0.37	[-0.93, 0.48]	3043
perceived comprehension, perceived comprehension x semantic centrality	-0.09	0.44	[-0.84, 0.77]	18531
semantic centrality, perceived comprehension x semantic centrality	-0.03	0.44	[-0.82, 0.80]	14118

Note: Population-level effect estimates are presented in logits. Rhat values for all parameters = 1.00. CI = credible interval. Eff Sample = number of effective samples, obtained using the bayestestR package (Makowski et al., 2019). QRI = Qualitative Reading Inventory. HLVA = Health Literacy Vocabulary Assessment. RBSOCF3 = Reader-Based Standards of Coherence Factor 3.

[a]Estimate refers to the mean of the marginal posterior distribution of the parameter. [b]Error refers to the standard deviation of the marginal posterior distribution of the parameter. [c]Credible intervals represent the upper and lower values within which 95% of the estimated parameter values in the posterior distribution are contained.

The estimated population-level effect of perceived comprehension on the probability of answering a comprehension question, targeting either semantically central or peripheral information, was similar to that obtained from model (9). An increase of 0.07 in the log odds of a correct response, on average, per unit increase in perceived comprehension was estimated for semantically central comprehension questions. The 95% credible interval for this effect included a small number of negative values (see Table 4.3). However, the majority of probability density in the posterior distribution was located over positive values (80% HDI = [0.02, 0.12]).

For comprehension questions assessing semantically peripheral information, the posterior mean for this interaction effect indicated an expected increase of 0.03 in the log odds per unit increase in perceived comprehension. This was a slightly higher estimated change compared to the estimates obtained from model (9). Despite this, there was limited change in the confidence in the direction of the effect, with considerable posterior plausibility associated with negative values (80% HDI = [-0.03, 0.09]). Participant-related variability in the predictive relationship, for both semantically central and peripheral comprehension questions, was equivalent to that estimated in model (9).

The main effects of the QRI, HLVA, and RBSOCF3 were estimated to be positive, though the 95% credible interval (see Table 4.3) and 80% HDI for the RBSOCF3 effect overlapped with zero (80% HDI = [0.00, 0.11]). This indicates that performance on the QRI and HLVA and, less certainly, the RBSOCF3, show a positive predictive relationship with response accuracy. For interpretation, predictions of the probability of a correct response, while holding all other predictors at their mean, can be calculated given the lowest and highest scores observed on each measure (QRI: 4 and 19; HLVA: 2 and 14; RBSOCF3: 23 and 63). For semantically central comprehension questions, given the lowest score observed in the data on the QRI, HLVA and RBSOCF3, the model-fitted predictions of observing a correct response are 0.53 (*SE* = 0.08), 0.47 (*SE* = 0.08), and 0.61 (*SE* = 0.08), respectively. Given the highest scores on these measures, these probabilities are 0.73 (*SE* = 0.06), 0.78 (*SE* = 0.05), and 0.69 (*SE* = 0.07). For semantically peripheral comprehension questions, given the lowest observed scores on the QRI, HLVA and RBSOCF3, the predicted probabilities of observing a correct response are 0.48 (*SE* = 0.08), 0.42 (*SE* = 0.08) and 0.56 (*SE* = 0.08). Given the highest observed scores, these probabilities are 0.68 (*SE* = 0.07), 0.75 (*SE* = 0.06), and 0.64 (*SE* = 0.07), respectively.

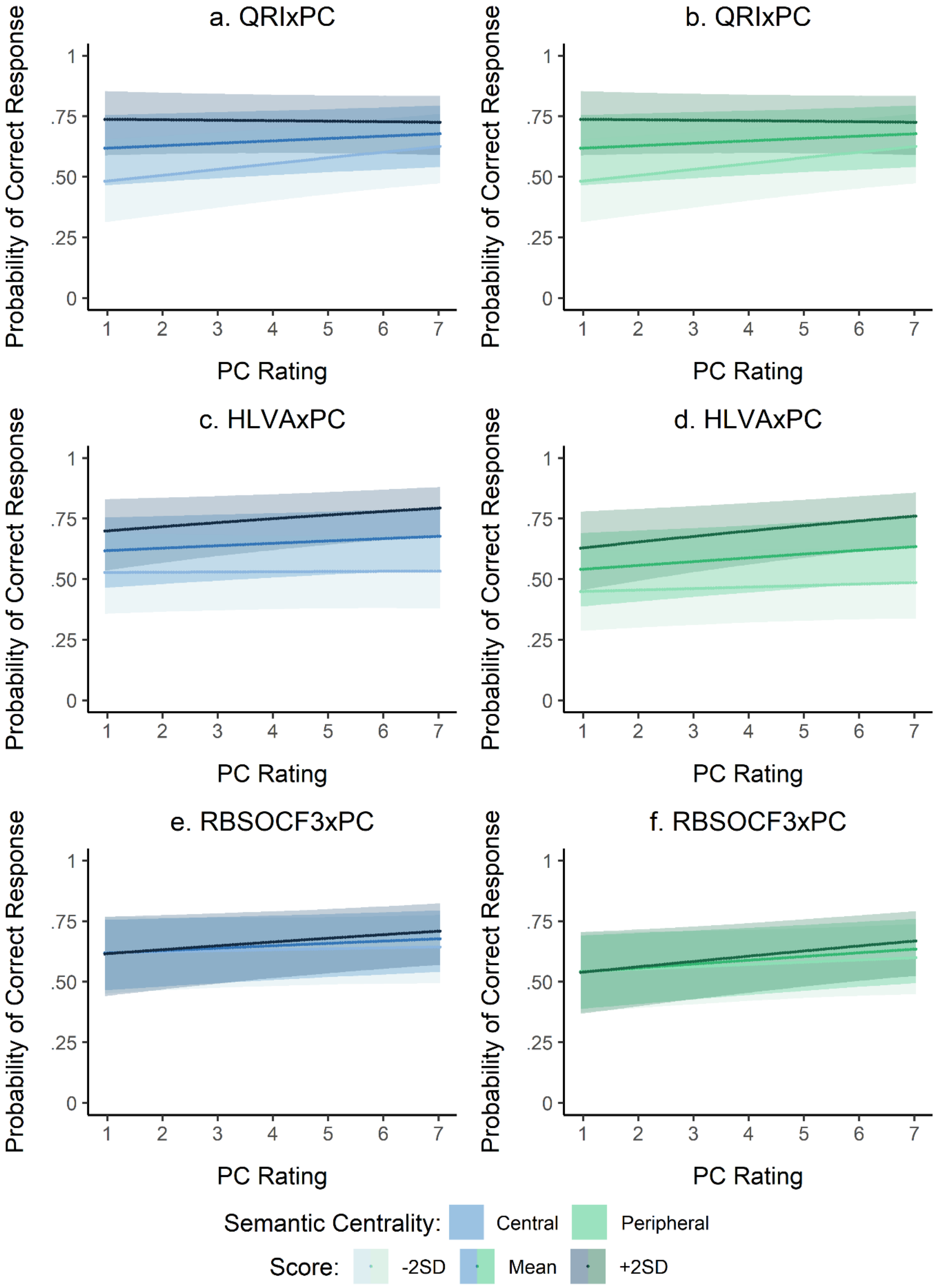
Limited evidence in support of an interaction between perceived comprehension and the QRI score was found. The posterior mean for this effect was negative and, although the upper bound of the 95% credible interval was positive (0.004), the 80% HDI was located across entirely negative values (80% HDI = [-0.07, -0.01]). An interaction of this direction would indicate that perceived comprehension judgements from individuals who score more highly on the QRI are more weakly associated with comprehension performance than judgements from individuals scoring lower on the QRI. This relationship is illustrated in Figure 4.8, showing the model-fitted predictions of the effect of perceived comprehension across varying levels of QRI performance, for semantically central (4.8a) and peripheral (4.8b) comprehension questions, respectively.

Little evidence in support an interaction was found between perceived comprehension and either the HLVA or RBSOCF3. The posterior mean for these interactions were positive, with the majority of probability density in the posteriors located over positive values. However, for both interactions, the 95% credible intervals (see Table 4.3) and the 80% HDIs included zero (‘perceived comprehension x HLVA’ 80% HDI = [0.00, 0.06]; ‘perceived comprehension x RBSOCF3’ 80% HDI = [-0.01, 0.05]). Model-fitted predictions of the expected effects are illustrated in Figure 4.8c to 4.8f, across varying levels of performance on the HLVA and RBSOCF3, for semantically central and peripheral comprehension outcomes. The magnitude of the predictive relationship between perceived and assessed comprehension can be seen to vary marginally across performance on the HLVA (Figure 4.8c and 4.8d) and RBSOCF3 (Figure 4.8e and 4.8f).

Estimates of group-level variability in the overall probability of answering a comprehension question correctly across participants, texts and questions remained largely comparable with the previous model (9). Between-question variance was the largest source of variability, followed by texts and participants. Inclusion of QRI, HLVA and RBSOCF3 scores in the model had a small impact on participant-level variability in response accuracy, indicating that these measures accounted for some, but not most, of the estimated variability between individuals. The considerable posterior uncertainty for covariance parameters was replicated in model (10).

Figure 4.8

Estimated Effect of Perceived Comprehension, by Semantic Centrality and Performance on the QRI, HLVA and RBSOCF3, in Study 2



Note. Marginal model-fitted predictions at the population-level, across varying levels of performance on the QRI (a. and b.), HLVA (c. and d.), and RBSOCF3 (e. and f.), are plotted, coloured according to whether the correct response corresponds to semantically central (blue) or peripheral (green) information. The panels show the expected probability of a correct response for the 'average' participant, responding to the 'average' question concerning an 'average' text, conditioned on the score plotted, with the 95% credible interval shaded. Scores of -2SD and +2SD refer to scores two standard deviations below and above the mean score, respectively. PC = perceived comprehension, QRI = Qualitative Reading Inventory, HLVA = Health Literacy Vocabulary Assessment, RBSOCF3 = Reader-Based Standards of Coherence Factor 3.

4.3.3 Sensitivity Analysis

A sensitivity analysis was conducted to explore the robustness of the findings of the main analysis (section 4.3.2), to alternative analytical choices. In this section, the alternative choices examined are first discussed. An outline of the motivation is then followed by an evaluation of the impact on the estimated effects of interest and a comparison of the predictive accuracy of the model specifications.

Alternative Model Specifications. Seven analytical choices were identified as potentially impacting the findings of Study 2: i) the measurement of semantic centrality of the information targeted by the comprehension questions, ii) the assumed shape of the predictive association between perceived and assessed comprehension, iii) the impact of multiple-choice questions on the question response process, iv) the chosen wording of the judgement prompt, v) the selection of priors for the effects of interest, vi) the use of participant exclusion criteria, and vii) the potential moderation of the relationship between judgements of comprehension and individual difference variables according to the semantic centrality of the comprehension question (i.e., three-way interactions). Each of these are discussed below. To limit repetition of content, where the motivation and implementation of an alternative specification is equivalent to that in Study 1, a brief description is given with reference to the fuller, previous discussion.

Measure of Semantic Centrality. Attempting to address issues concerning the existing definitions and methods of measuring centrality, a procedure was developed in Study 2 to identify semantically central and peripheral information within texts. While this approach was successfully applied, there were a number of shortcomings. Most importantly, the extent to which this approach was able to provide a valid and reliable measure of semantic centrality was unclear. As recognised when applying the procedure, semantic centrality may be better characterised as a continuum of varying importance rather than as a binary variable. These issues are highly pertinent, given the importance of the quality of measurement needed to robustly address RQ3.

A small-scale study was conducted to evaluate how the identification approach compared to ratings of perceived semantic centrality, similar to previous research (e.g., Yeari et al., 2017). As described in the method section (see section 4.2.4), judgements of semantic centrality were obtained for each of the 78 idea units. Ratings of centrality showed a weak correspondence with the identification approach used in Study 2: idea units identified as central were rated as more important, on average, than those identified as peripheral (see Figure 4.3). However, considerable variability was observed, suggesting that these two measures do not exclusively capture the same underlying variable.

To evaluate whether the findings of Study 2 are robust to the apparent variability in measurement, the judgements of perceived semantic centrality were used in place of the binary indicator variable to provide an alternative measure of semantic centrality. To include rated centrality as a predictor, the average rating of semantic centrality for each idea unit was first calculated. Given the small sample size of centrality judgements, the average rating was specified within the model as a variable associated with uncertainty. Therefore, the models defined in (9) and (10) were respecified as measurement error models (or ‘errors-in-variables’ models).

Measurement error models are used to specify that observations of a variable are a combination of the true underlying value in addition to some variance (or error) associated with the measurement process. The simplest form of measurement error model assumes that the variance introduced through measurement is normally distributed. In the context of rated centrality, it may be reasonable to assume that there is a true, underlying rated value of centrality μ_{true} and that variance in observed ratings σ corresponds to normally distributed individual differences in the perception of centrality. This may be expressed as:

$$Rated\ Centrality \sim Normal(\mu_{true}, \sigma)\ .$$

Based on the assumption that variability in ratings of perceived centrality is normally distributed, the sampling distribution for μ_{true} can be defined as:

$$\mu_{rated} \sim Normal\left(\mu_{true}, \frac{\sigma}{\sqrt{N}}\right)\ .$$

Where *N* is the number of observed ratings. The average of a sample of ratings (μ_{rated}) will tend toward the true centrality rating as *N* approaches infinity. The spread of the distribution σ of individual variability for each item is unknown. However, the standard deviation of the sample may be used as an estimate of the true standard deviation. Using the average rated centrality calculated for each item, the measurement error in the model can be specified using the standard error of the estimated mean.

For the models defined in (9) and (10), observations of $Centrality_k$ were replaced with the average rated semantic centrality for each comprehension question, associated with uncertainty (the standard error of the mean), using the measurement error model syntax in brms package (Bürkner, 2017, 2019). It is important to note, however, that judgements of centrality are ordinal-level data and variance around the true underlying value of rated centrality, should it exist, is unlikely to be normally distributed at the bounds of the rating scale. Therefore, the extent to which these model specifications yield valid estimates is unclear. However, this model provides a method to evaluate an alternative measure of centrality whilst incorporating uncertainty in this measure, albeit according to simplifying assumptions which may only approximate reality.

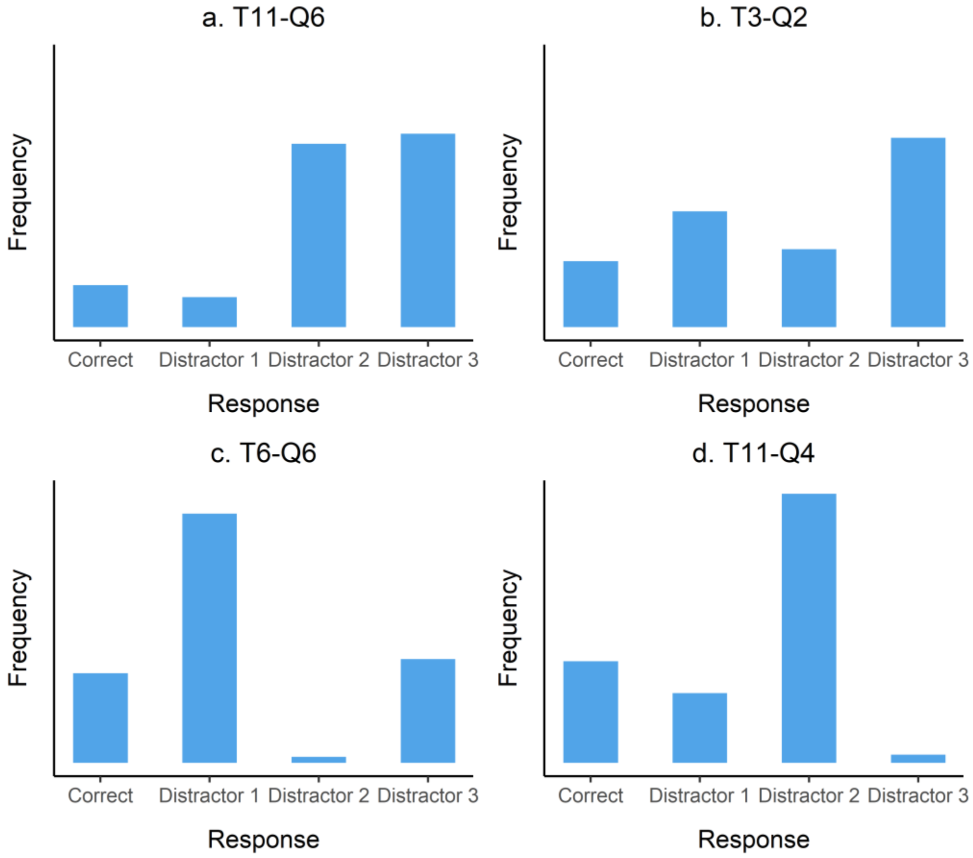
Nonlinear Relationship. The models defined in (9) and (10) assumed that the relationship between perceived and assessed comprehension is linear on the logit scale and that judgements of perceived comprehension could be treated as interval-level data. As discussed in the sensitivity analysis in Study 1 (Chapter 3; section 3.3.3), such an assumption, if incorrect, can lead to inaccurate estimates for the predictive relationship. Instead, estimating the monotonic association, as described by Bürkner and Charpentier (2020), may arguably better capture the ordinal nature of metacomprehension judgements. The models defined in (9) and (10) were altered, therefore, to estimate the monotonic effect of perceived comprehension and associated interactions. Six simplex parameters (see Chapter 3; section 3.3.3) were estimated, capturing the expected difference in the probability of a correct response between adjacent ratings of comprehension.

Random Guessing. In the metacomprehension task, each of the six comprehension questions per text (three targeting semantically central information and three targeting semantically peripheral information) were presented with multiple response options. As discussed in the sensitivity analysis in Study 1 (Chapter 3; section 3.3.3), questions with closed-form responses can introduce bias in the estimation of the predictive relationship between perceived and assessed comprehension (Vuorre & Metcalfe, 2022). A model in which the intercept is constrained can be used to incorporate a random guessing response process, estimating the influence of the covariates in addition to some fixed probability of a correct response due to chance alone (Bürkner, 2022).

Although this model offers an effective method to model the probability of correct responses which can be attributed to random guessing, the assumption that individuals engage in this response process may not be accurate (Higham, 2007; Vuorre & Metcalfe, 2022). Examining the selection of response options to comprehension questions which have the lowest levels of response accuracy permits a limited evaluation of this assumption. As can be seen in Figure 4.9, showing responses to four questions with the lowest response accuracy rates, the observed probabilities of selecting the incorrect response options appears inconsistent with a random selection process.

Figure 4.9

Responses to Four Comprehension Questions with the Lowest Response Accuracy in Study 2



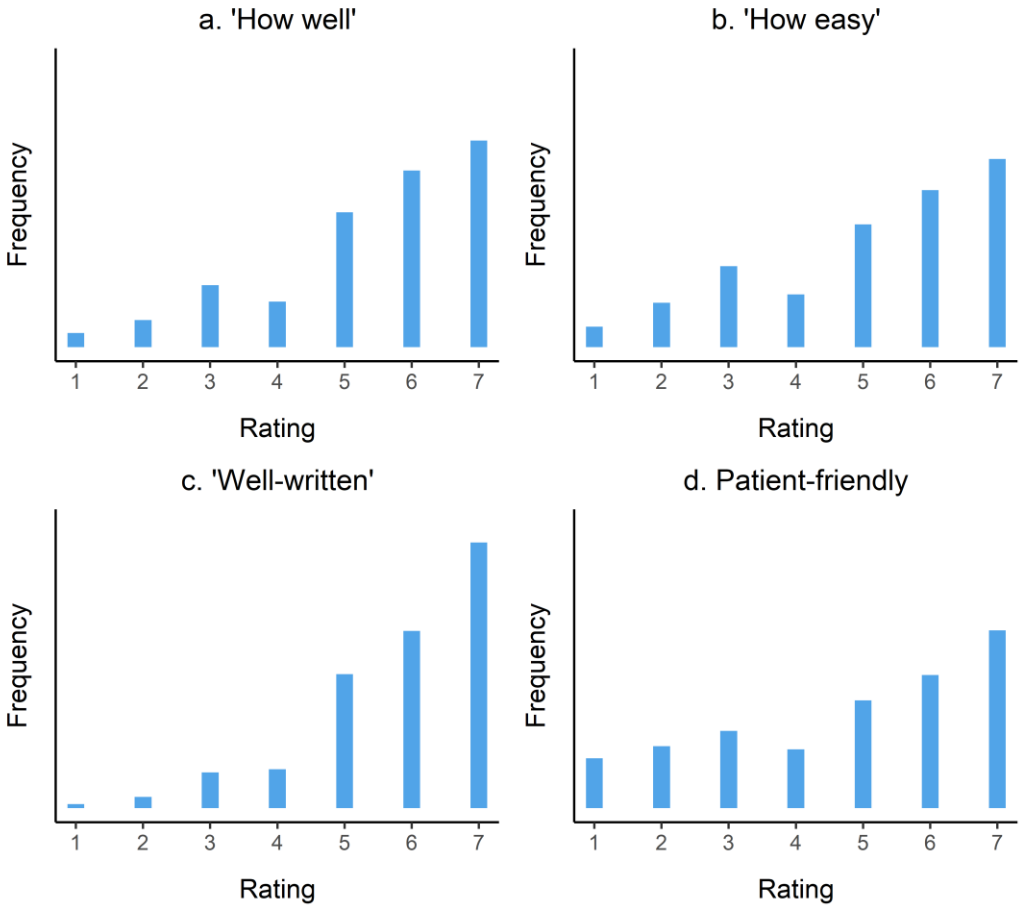
Note. T = text. Q = question.

Nevertheless, this does not exclude the possibility that participants engage in guessing to some extent alongside some form of knowledge-driven selection procedures (e.g., Embretson & Weztel, 1987). While the role of random guessing in the response process is unclear, statistically modelling a random guessing response process provides a method to evaluate the bias that is introduced in the estimated relationship between perceived and assessed comprehension if this assumption is true but not accounted for. The models defined in (9) and (10), therefore, were also altered to incorporate a random guessing response process.

Alternative Prompts. Judgements were elicited using four additional prompts, due to potential sensitivity of the relationship between perceived and assessed comprehension to the wording of the prompt (Pilegard & Mayer, 2015). The distributions of responses to the alternative prompts are shown in Figure 4.10.

Figure 4.10

Distributions of Judgements in Response to the Four Alternative Prompts in Study 2



Note. ‘How well’ = ‘Overall, how well do you understand the text?’. ‘How easy’ = ‘How easy was it to understand the text?’. ‘Well-written’ = ‘How well was the text written? (In terms of spelling, punctuation and grammar)’. ‘Patient-friendly’ = ‘How patient-friendly was the text? (In terms of technical or medical words)’.

As described in the sensitivity analysis of Study 1 (Chapter 3, section 3.3.3), the models defined in (9) and (10) were refit using observed responses to the alternative judgement prompts as the predictor $Judgement_{ij}$ to examine the possible influence of prompt variation.

Variation in Priors. Weakly informative priors were chosen for model parameters in the main analysis (section 4.3.2). The extent to which the posterior distributions for the effect parameters are robust to reasonable variation in the prior specification for effect parameters was examined to evaluate the impact of the choice of priors. The models defined in (9) and (10) were refit with looser and tighter priors, as described in the sensitivity analysis of Study 1 (Chapter 3; section 3.3.3).

No Exclusions. Participants were excluded from the main analysis if any of their reading times for the 13 health-related texts were faster than a reading rate of 300 wpm. The motivation for excluding participants and the potential issues with a binary decision threshold for exclusion are discussed in the sensitivity analysis of Study 1 (Chapter 3; section 3.3.3). To evaluate whether the estimated effects were influenced by participant exclusion, the models defined in (9) and (10) were refitted using the full dataset without exclusions (*I* = 300).

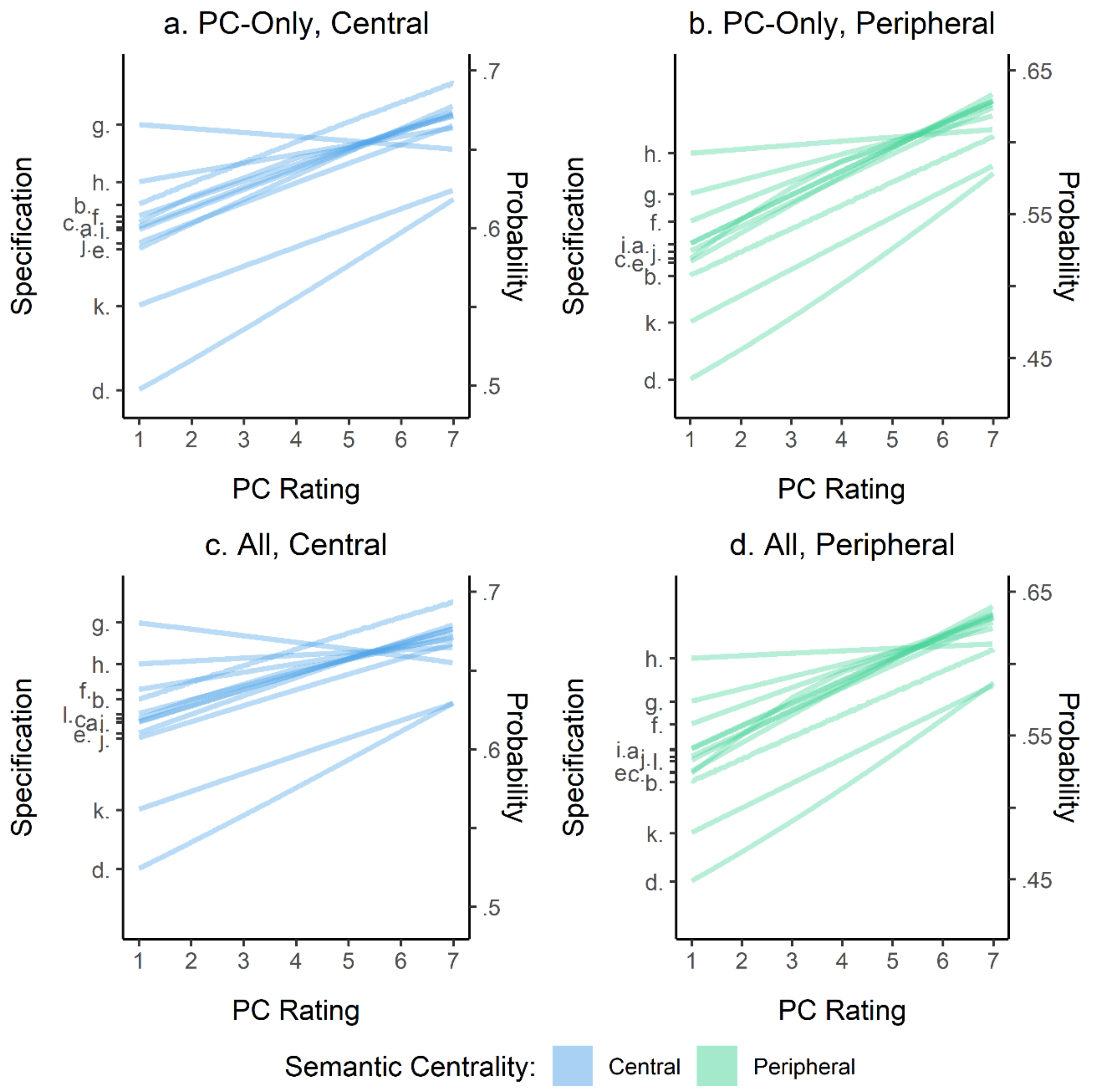
Three-Way Interactions. The model defined in (10) included interactions between the three individual difference measures (QRI, HLVA, and RBSOCF3) and judgements of perceived comprehension. This permitted the magnitude of the predictive relationship between perceived and assessed comprehension to vary according to an individual’s reading ability, background knowledge and reader-based standards of coherence. However, this model specification did not permit these relationships to vary with respect to the semantic centrality of the information targeted in the comprehension question.

It could be suggested that three-way interactions between metacomprehension judgements, semantic centrality, and the individual difference measures may be reasonable. For example, individuals with lower reading ability may have difficulties in identifying information which is semantically central in the macrostructure of the text, leaving them less able to evaluate the quality of this information when forming judgements of comprehension. Consequently, their metacomprehension judgements may be less predictive of correctly responding to questions concerning semantically central information than individuals with greater reading ability. Failing to account for such relationships may produce bias in the estimates reported in the main analysis. To examine whether such dependencies may be observed, therefore, the model defined in (10) was extended to include three-way interactions between judgements of perceived comprehension, semantic centrality, and the three individual difference measures.

Model Comparison: Effects of Interest. For models which included only metacomprehension judgements, semantic centrality, and the interaction between these variables as population-level predictors, as in (9), the estimated effects were generally highly similar across the alternative model specifications considered. The posterior means for the effect of judgements on the probability of answering a semantically central comprehension question correctly remained positive in all but one model. The posterior means for the change in the log odds, given a semantically peripheral question, remained positive in all but two model specifications. Limited participant-level variability in these effects was consistently replicated. Figure 4.11 shows the model-fitted predictions of the expected probability of answering either a semantically central (4.11a) or semantically peripheral (4.11b) comprehension question correctly. Note that the range of the vertical axis is reduced in this figure to more clearly show the differences between estimated effects.

Figure 4.11

Estimated Effect of Perceived Comprehension for Each Model Specification in Study 2



Note. Marginal model-fitted predictions at the population-level for model specifications including only participants' judgements as the predictor for semantically central (a.) and peripheral (b.) questions, and all predictors for semantically central (c.) and peripheral (d.) comprehension questions are plotted. Model specifications a-l refer to: a. main analysis model, b. measurement error, c. monotonic effect, d. constrained intercept, e. prompt 'Overall, how well do you understand the text?', f. prompt 'How easy was it to understand the text?', g. prompt 'How well was the text written? (In terms of spelling, punctuation and grammar)’, h. prompt ‘How patient-friendly was the text? (In terms of technical or medical words)’, i. looser effect priors, j. tighter effect priors, k. no exclusion criteria, l. three-way interactions. PC = perceived comprehension.

For both semantically central and peripheral questions, the greatest similarity to the effect estimates obtained from model (9) were observed in models which used looser prior specifications for population-level effects (labelled as specification i. in Figure 4.11a and 4.11b). High similarity was also observed in models which assumed a monotonic relationship between perceived and assessed comprehension, used the alternative judgement prompts 'Overall, how well do you understand the text' and 'How easy was is to understand the text', and used tighter prior distributions (labelled as specifications c., e., f., and j., respectively). In addition, despite the lower estimated intercept, when no exclusion criteria were applied (labelled as specification k.), the estimated relationship between judgements and performance were equivalent to that estimated in the main analysis (labelled as specification a.). Likewise, incorporating a random guessing response process (labelled as specification d.), resulted in a lower estimated intercept. However, the posterior means for the effects of judgements were estimated to be slightly larger in this model.

The largest differences in the estimated effects were found for models which used alternative judgements prompts which concerned the quality of the writing. However, the impact of these prompts was dependent on whether the comprehension question assessed semantically central or peripheral information. For central comprehension questions, one model specification produced a negative estimate for the effect of judgements. Labelled as specification g. in Figure 4.11a, this model used the alternative judgement prompt ‘How well was the text written? (In terms of spelling, punctuation and grammar)’. This was the only specification to estimated negative individual-level predictions for the relationship between perceived and assessed comprehension.

With respect to peripheral comprehension questions, the largest difference between model specifications was observed in the model using the alternative judgement prompt 'How patient-friendly was the text? (In terms of technical or medical words)’. This model estimated a negative coefficient for the change (interaction) in effect of judgements (labelled specification h.). Similarly, a negative estimate for the interaction effect was observed when rated semantic centrality was used as the predictor (measurement error model, labelled as specification b.). A negative estimate for the change (interaction) in the log odds of a correct response given a peripheral comprehension question indicates a stronger relationship between judgements and correct responses to semantically central comprehension questions, compared to semantically peripheral comprehension questions. However, similar to the main analysis, the 95% credible interval for the estimates from these models overlapped with zero.

For models which included QRI scores, HLVA scores, RBSOCF3 scales, and the interactions between these variables as population-level predictors, as in (10), estimates were predominantly similar for the main effects across the alternative model specifications. Variation in the estimated relationship between perceived and assessed comprehension of semantically central and peripheral information was equivalent to that described above. This similarity can be seen in Figure 4.11, showing the model-fitted predictions of the expected probability of answering a semantically central (4.11c) or peripheral (4.11d) comprehension question correctly. Estimated posterior means for the effect of judgements, in addition to participant-level estimates of the association, remained positive in all but one model specification. For semantically central comprehension questions, using the alternative judgement prompt 'How well was the text written? (In terms of spelling, punctuation and grammar)' (labelled as specification g.) produced a negative posterior mean.

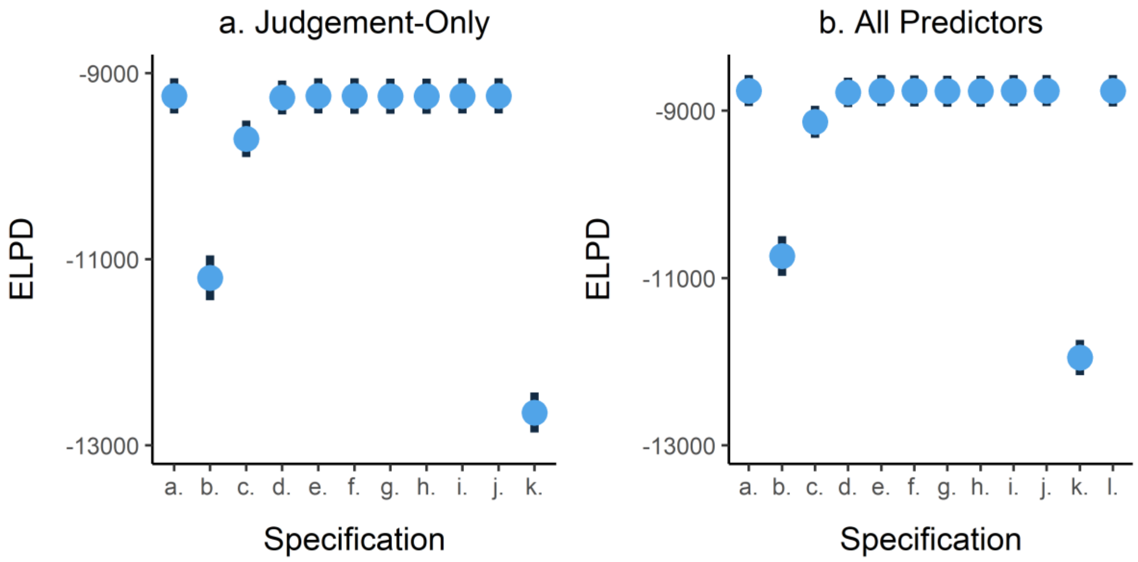
The main effects of QRI score, HLVA score and RBSOCF3 score were consistently estimated to be positive across alternative specifications. Not applying the exclusion criteria resulted in higher magnitudes for these effects, particularly QRI score and RBSOCF3 score, with the 95% credible interval for the latter located entirely over positive values (95% CI = [0.02, 0.18]). In contrast, with respect to the interactions between these variables and perceived comprehension, greater variability in the magnitude and direction was estimated across alternative specifications. Four models indicated reduced support for interactions effects, whereas two models indicated increased support. Modelling three of the alternative prompts as predictors of assessed comprehension, including ‘How easy was it to understand the text?’, ‘How well was the text written? (In terms of spelling, punctuation and grammar)’, and ‘How patient-friendly was the text? (In terms of technical or medical words)’, or not employing the exclusion criteria resulted in the estimated posterior means shifting towards zero for the interaction effects of QRI, HLVA and RBSOCF3 scores. In contrast, modelling a random-guess response process yielded a 95% credible interval for the interaction effect of QRI scores which was located entirely over negative values (95% CI = [-0.04, -0.01]). In addition, including three-way interactions in the model produced robust support for an interaction between RBSOCF3 score and semantic centrality. Scoring more highly on the RBSOCF3 was estimated to increase the probability of a correct response on semantically central comprehension questions (95% CI = [0.02, 0.21]), while reducing the probability of a correct response on semantically peripheral questions (95% CI = [-0.21, -0.05]).

Model Comparison: Predictive Accuracy. A comparison of the relative predictive accuracy of the model specifications was conducted to evaluate whether predictive performance differed across the models. As in Chapter 3 (section 3.3.3), predictive performance was evaluated in two ways for each model: i) calculating the estimated expected log predictive density (ELPD) and ii) posterior predictive checks. The ELPD was calculated using the Widely Applicable Information Criterion (WAIC) computation (Vehtari et al., 2016; Watanabe, 2010), to provide a measure of out-of-sample predictive accuracy. Posterior predictive accuracy was evaluated by calculating the discrepancy between observed response accuracy and model predictions, to provide a measure of within-sample predictive accuracy (Gabry et al., 2019).

With respect to out-of-sample predictive accuracy, Figure 4.12 shows the estimated ELPDs for the each of the alternative specifications, for models omitting (4.12a) or including (4.12b) the individual difference measures (QRI, HLVA, RBSOCF3 scores and associated interactions). In this figure, the points show the estimated ELPD for each model and the horizontal bars show the estimated uncertainty (three times the standard error of the estimated ELPD). Note that direct comparisons between estimated ELPDs from models fitted to different sized datasets (here, specification k., referring to models fitted with no exclusion criteria) are not appropriate, given that possible differences in predictive accuracy are indistinguishable from the impact of additional observations.

Figure 4.12

ELPD Estimates for Each Model Specification in Study 2



Note. Points (blue circles) show the estimated ELPD and bars show the standard error of the estimate, for model specifications including (a.) only participants' judgements as the predictor and (b.) all predictors. Model specifications a-l refer to: a. main analysis model, b. measurement error, c.

monotonic effect, d. constrained intercept, e. prompt ‘Overall, how well do you understand the text?’, f. prompt ‘How easy was it to understand the text?’, g. prompt ‘How well was the text written? (In terms of spelling, punctuation and grammar)’, h. prompt ‘How patient-friendly was the text? (In terms of technical or medical words)’, i. looser effect priors, j. tighter effect priors, k. no exclusion criteria, l. three-way interactions. ELPD = expected log predictive density.

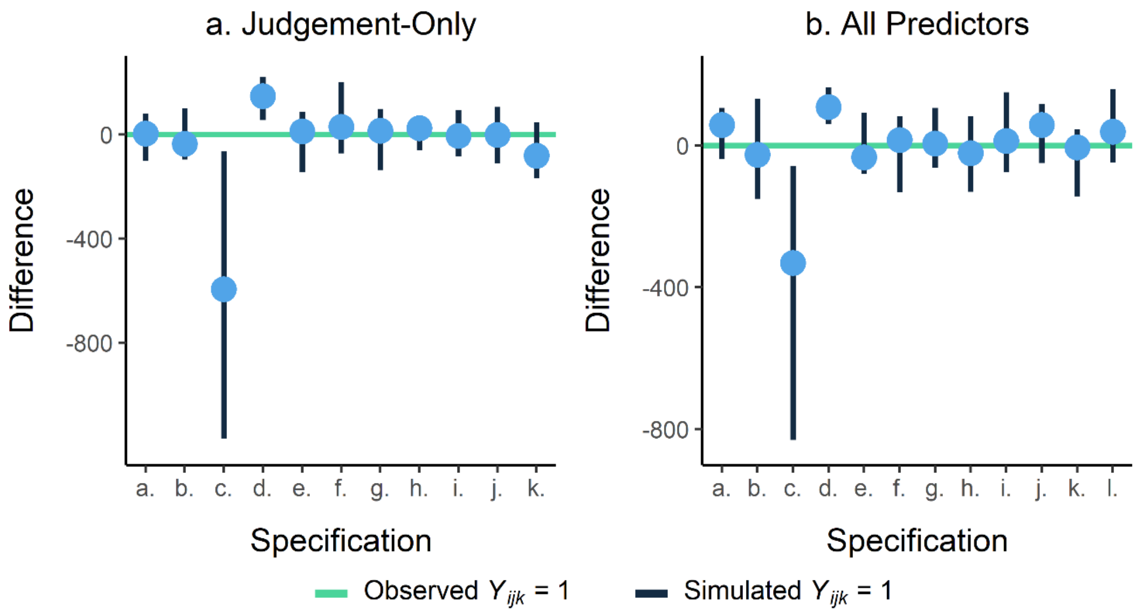
Minimal differences in estimated ELPD were found for the majority of model specifications. However, lower predictive accuracy was obtained when rated semantic centrality was used in place of the binary classification variable (measurement error model, labelled as specification b.). Modelling the relationship between perceived and assessed comprehension as monotonic also produced a slightly lower estimated ELPD when individual difference measures were not included as predictors (labelled as specification c. in Figure 4.12a).

With respect to within-sample predictive accuracy, Figure 4.13 shows the estimated posterior predictive accuracy for each of the different model specifications, for models omitting (4.13a) or including (4.13b) the individual difference measures (QRI, HLVA, and RBSOCF3 scores). In this figure, the light blue circles show the mean difference between observed and simulated correct responses and the dark blue lines show the range of differences estimated.

In the majority of model specifications, the difference between observed and simulated response accuracy appeared limited and unbiased. However, posterior predictions from models which assumed a monotonic relationship between perceived and assessed comprehension (labelled as specification c.) tended to substantially underestimate correct responses. Posterior predictions from models with a constrained intercept (labelled as specification d.) tended to slightly overestimate correct responses.

Figure 4.13

Posterior Predictive Accuracy for Each Model Specification in Study 2



Note. Plotted according to model specifications including (a.) only participants' judgements as the predictor and (b.) all predictors. Light blue circles show the mean difference between observed correct responses ($Y_{ijk} = 1$) and simulated correct responses. Dark blue lines show the range of differences estimated between observed and simulated correct responses. Model specifications a-l refer to: a. main analysis model, b. measurement error, c. monotonic effect, d. constrained intercept, e. prompt 'Overall, how well do you understand the text?', f. prompt 'How easy was it to understand the text?', g. prompt 'How well was the text written? (In terms of spelling, punctuation and grammar)', h. prompt 'How patient-friendly was the text? (In terms of technical or medical words)', i. looser effect priors, j. tighter effect priors, k. no exclusion criteria, l. three-way interactions.

4.4 Discussion

Study 2 was conducted to address RQ3 and RQ4 (see section 4.1.2). In this section, the main findings are discussed with respect to existing research, followed by a consideration of the theoretical implications of the findings and the limitations of this research.

4.4.1 Main Findings

A weakly positive relationship between perceived and assessed comprehension was found in the present study. This relationship did not reliably vary according to the semantic centrality of the information assessed by the question. The relationship was estimated to be positive in all but one specification considered in the sensitivity analysis. Responses to the prompt ‘How well was the text written? (In terms of spelling, punctuation and grammar) showed a negative predictive association with response accuracy on semantically central comprehension questions. Limited individual variability in the relationship was estimated across all model specifications. Overall, therefore, these findings suggest that higher judgements of comprehension correspond to a greater likelihood of demonstrating evidence of understanding, with this relationship not reliably influenced by the position of the assessed information within the macrostructure constructed from a text. While readers show limited predictive accuracy in discriminating between texts which are more or less well understood, these judgements positively associate with comprehension of both semantically central and peripheral information across individuals.

The limited predictive relationship between metacomprehension judgements and assessed comprehension, regardless of the semantic centrality of the information targeted in the comprehension questions, is consistent with the findings of previous research (Dunlosky & Lipko, 2007; Lin & Zabrucky, 1998; Maki, 1998; Prinz et al., 2020a; Yang et al., 2022). The limitations of the analytic approaches taken in these previous studies aside (see Chapter 2), the comparability between the results obtained in Study 2 and previous research further supports the argument that measures of perceived and assessed comprehension show a weakly positive association, on average. In addition, the direction and magnitude of the estimated effects observed in Study 2 are consistent with those observed in Study 1, indicating that the relationship is robust to the alterations in the method of constructing questions and the range of the judgement response scale.

Contrary to previous research, however, variability in the relationship between perceived assessed comprehension was estimated to be limited in the present study. Previously, a wide range of individual-level metacomprehension accuracy estimates have been reported, with remarkably high or negative associations estimated for some individuals (Chiang et al., 2010; Glenberg & Epstein, 1985; Jee et al., 2006). Such variability in the magnitude of the predictive relationship was not observed in the present study or in Study 1. In Study 2, negative associations were estimated only when considering the relationship between judgements concerned writing quality and performance on semantically central comprehension questions. Given the divergence between previous research and the findings of both Study 1 and 2, previously reported large variability in metacomprehension likely reflects an exaggeration of individual-level estimates, due to an insufficient quantity of observations to resolve the estimates with accuracy and precision, under typically adopted analytic approaches (Schönbrodt & Perugini, 2013; Yarkoni, 2009).

The findings of Study 2 also provide insight into the relationship between judgements of comprehension and the demands of the comprehension measure. Previously, in assessing understanding of the text, Griffin, Wiley and Thiede (2019) and Pilegard and Mayer (2015) presented participants with memory-for-detail and inference-based questions, with the latter requiring a greater integration of information than the former. Similarly, in Study 2, semantically central questions likely required a greater amount of integration of information than semantically peripheral questions, which typically assessed details stated in the text. While Pilegard and Mayer (2015) found no reliable differences in metacomprehension accuracy between question types, Griffin, Wiley and Thiede (2019) found that judgements were more predictive of performance on memory-for-detail questions than inference-based questions. In Study 2, the direction of the posterior mean of the interaction effect was consistent with the results of Griffin, Wiley and Thiede (2019), indicating some support for a stronger predictive relationship between judgements and questions which require less integration of information. However, consistent with Pilegard and Mayer (2015), this effect was not reliable. The results of Study 2, therefore, provide converging evidence that metacomprehension accuracy may not be influenced by the integrative demand of comprehension questions, though further research may yield evidence for a positive effect of decreasing processing demands.

The analyses of how individual differences in reading ability, background knowledge, and reader-based standards of coherence may influence the predictive relationship between perceived and assessed comprehension failed to reveal any reliable interactions. Model specifications considered in the sensitivity analysis indicated that evidence in support of such interactions may be sensitive to analytic choices. The selection of the judgement prompt, specification of interactions further moderated by semantic centrality, and the use of exclusion criteria were features which produced notable differences in interaction estimates. In the majority of models considered, however, a considerable portion of the posterior distribution for the interaction between reading ability and perceived comprehension judgements was located over values below zero. In direct contrast to the findings of Study 1, this indicates support for a small negative relationship between these variables. The strongest evidence for this interaction was observed when the model intercept was constrained to incorporate a random guessing response process.

Previous research which has considered the relationship between metacomprehension accuracy and reading ability has provided mixed results, similar to the inconsistencies observed between Study 1 and 2. It has been suggested that there may be no relationship between metacomprehension accuracy and reading ability (Lin et al., 2000; Maki et al., 2005), or that a significant positive association between these variables exists (Griffin et al., 2008). The support for a negative association observed in Study 2 would mean that the judgements of individuals who score lower on measures of reading ability are more predictive of comprehension performance than those made by individuals with higher reading ability scores. Such a relationship would contradict suggestions that lower ability readers may struggle to simultaneously monitor the quality of their comprehension while engaging in comprehension-building processes (Griffin et al., 2008).

The shifting evidence for a relationship between reading ability and metacomprehension accuracy, in both previous research and in Study 1 and 2, suggests that this association may be influenced by differences between studies. Given the comparability between Study 1 and 2, the apparent reversal in the direction of the estimated relationship between reading ability and perceived and assessed comprehension is unexpected. A potential explanation for this may be suggested, related to text readability. Research suggests that the relationship between reading ability and metacomprehension accuracy may depend on text difficulty (Maki et al., 2005; Weaver & Bryant, 1995). In Study 1, stimulus texts were similar in Flesch Reading Score (Flesch, 1948) readability, whereas Study 2 featured a number of more difficult texts. The different pattern of results observed in Study 1 and 2 could, therefore, be consistent with suggestions of a complex relationship between text characteristics, reading ability and metacomprehension accuracy. However, as several factors may differ between studies, further research is required identify the circumstances which may produce an association between these variables.

With regards to the relationship between background knowledge and metacomprehension accuracy, consistent with previous research (Jee et al., 2006; Griffin et al., 2009; Shanks & Serra, 2014) and the results of Study 1, no support for an interaction between these variables was observed. A lack of support for an association between background knowledge and judgements of comprehension was replicated in all models in the sensitivity analysis. This study, therefore, adds to the existing body of evidence suggesting that variation in background knowledge does not influence an individual’s capacity to demonstrate predictive accuracy in their comprehension judgements. However, this does not preclude an effect of background knowledge on measures of absolute accuracy. Background knowledge may influence measures of metacomprehension accuracy calculated using difference scores, with greater expertise reducing the bias in judgements (Griffin et al., 2009; Jee et al., 2006).

A lack of support for an interaction between perceived comprehension and reader-based standards of coherence was also observed in Study 2. This indicated that self-reported perceptions of willingness to engage in comprehension-building processes to achieve a desired quality of understanding did not influence the relationship between perceived and assessed comprehension. In addition, a three-way interaction featuring semantic centrality was explored in the sensitivity analysis. Such an interaction was considered plausible since individuals may vary in terms of what they consider to be main ideas (Baker & Anderson, 1982; Brown & Smiley, 1977; Calloway, 2019; van den Broek et al., 2011; van den Broek & Helder, 2017), with individuals with high standards of coherence potentially more likely to produce more exhaustive macropropositional structures (Christianson, 2016; Ferreira et al., 2002). No support for a three-way interaction was found, however, an interaction between reader-based standards of coherence and semantically central comprehension questions was found. This suggests that a higher motivation to achieve comprehension increases the probability of demonstrating understanding of information which features centrally within the macrostructure of the text. This finding is consistent with the view that that a desire to achieve understanding promotes greater comprehension of information which requires effortful text-integration processes (Calloway, 2019).

Although not the focus of the present study, positive main effects of reading ability and background knowledge on the probability of correctly answering a comprehension question were observed in all models considered. This is consistent with the role of general comprehension-based skills and domain-specific knowledge specified in multiple models of reading comprehension (Ahmed et al., 2016, Cromley & Azevedo, 2007; Gough & Tunmer, 1986; Perfetti, 1999; Perfetti & Stafura, 2014; Tzeng et al., 2005; van den Broek & Helder, 2017). In addition, unit increases in both reading ability and background knowledge were found to lead to greater increases in the probability of demonstrating understanding than a unit increase in perceived comprehension, though the latter remained a robust predictor in the model when individual difference measures were included. This indicates that reading ability and background knowledge more strongly predict comprehension outcomes than judgements of comprehension, however judgements capture variance which is uniquely predictive of comprehension on a given text.

4.4.2 Theoretical Implications

The relationship between perceived and assessed comprehension evidenced in Study 2 indicates that metacomprehension judgements are weakly predictive of performance on comprehension questions which assess either semantically central or peripheral information in health-related texts, on average. As discussed in Chapter 2, given the limitations of available approaches to measurement, the estimates obtained in Study 2 should not be directly interpreted as a measure of individuals' capacity to evaluate the status of their understanding of text (Paulewicz et al., 2020). As such, strong conclusions regarding individuals' ability to internally discriminate between texts which are more or less well understood, or the basis of such judgements, are not justifiable. However, it has been previously shown that regression slopes are sensitive to differences in underlying discrimination accuracy (Rausch & Zehetleitner, 2017). The measurement approach adopted in Study 2, therefore, arguably provides some insight into underlying metacomprehension processes and the predictive relationship is not produced merely as an artefact of measurement.

The dominant view of metacomprehension accuracy suggests that the average, weakly positive alignment between measures of perceived and assessed comprehension occurs due to individuals utilising sources of information which are poorly diagnostic of comprehension when forming judgements of comprehension (Koriat, 1997; Thiede et al., 2019). Considerable between-individual variability in estimated metacomprehension accuracy is assumed to occur due to wide differences in the informativeness and suitability of the cues used (Thiede et al., 2010). However, this view is not consistent with the relative homogeneity between individuals in the predictive relationship observed in both Study 1 and 2. Instead, these findings cohere with accounts which propose a default tendency to rely on a primary source of information in forming judgements of comprehension across individuals. Yet, as discussed in Chapter 1 (section 3.4.2), previous suggestions of individuals utilising a primacy type of cue to inform judgements, including disruptions to text processing or a memory-based test heuristic (Dunlosky, Rawson & Hacker, 2002; Griffin, Wiley & Thiede, 2019), provide a limited account of the observations of Study 1 and 2.

An alternative primary source of information which may plausibly underlie metacomprehension judgements was explored in the present study, namely metacognitive experiences relating to the construction of coherent macropropositional structure of the text. It was proposed that limited individual variability in the predictive relationship corresponds to minimal variation in the contents of the macropropositional structure between skilled readers with a shared reading goal of understanding the text. Further, it was suggested that the overall weak predictive relationship reflected a lack of correspondence between the basis of metacomprehension judgements and the information assessed in the comprehension questions. No evidence which supported either of these suggestions was observed. Judgements were no more predictive of comprehension of semantically central information than semantically peripheral information. In fact, the analysis indicated limited support for the inverse relationship, such that judgements are more predictive of specific details within texts, though this effect was not reliable. Further, despite increasing the alignment between the hypothetical bases of perceived and assessed comprehension, the magnitude of the predictive relationship did not differ from that observed in Study 1. Overall, therefore, in so far as the estimated predictive relationship reflects underlying processes, it does not appear that judgements of comprehension are strongly informed by macropropositional coherence.

The conclusion that disruptions in achieving macropropositional coherence does not influence metacomprehension judgements may be somewhat premature, however, due to limitations with the measurement of comprehension. In designing the comprehension questions in Study 2, the identification of semantically central and peripheral information was carried out by two researchers (see Appendix J). While it is reasonable to expect that skilled readers' macrostructure of a text, given a shared goal of understanding, is likely to show some similarities (Kintsch, 1994; Kintsch & van Dijk, 1978; van den Broek et al., 2011; van Dijk & Kintsch, 1983), it is unclear how well the semantic networks constructed from the texts by two individuals reflect the macrostructure constructed by an average reader. Evidence of potential divergence was observed in the judgements of semantic centrality collected for each idea unit (see Figure 4.3), with a number of individuals reporting ratings of perceived importance to the overall meaning of the text which were incongruent with the centrality classification. This indicates that comprehension questions may not have effectively assessed comprehension of the core set of macropropositions and peripheral propositions. The lack of an observed effect, therefore, may be attributed to constraints on metacomprehension accuracy given the discrepancies between what was judged and what was assessed (Weaver, 1990; Wiley et al., 2005).

Nonetheless, while evidence was not observed to support the view that metacomprehension judgements are informed by macropropositional coherence, the findings of this study provide valuable insight into metacomprehension judgements. Arguably, there are two categories of factors which may produce the weak predictive relationship observed, relating to aspects which are either internal or external to the judgement process. A weak alignment due to internal factors, such as a reliance on a source of information which is suboptimal in evaluating comprehension, would imply that the internal evaluation of comprehension quality is largely inaccurate (Thiede et al., 2010). External factors, corresponding to the methods of measuring perceived and assessed comprehension, may produce a weak alignment by introducing measurement error (Weaver, 1990; Wiley et al., 2005). For example, individuals may have difficulties in translating an internal judgement state into an ordinal measurement. Likewise, limited variability in the predictive relationship arguably occurs due to factors which are internal and external to the judgement process between individuals. For example, the internal judgement process may be influenced by the use of varying sources of information, across individuals, to supplement judgements (e.g., Linderholm et al., 2008). The observed judgement, in contrast, may be influenced by factors such as differences in willingness to report a lack of understanding.

Whether the cause of the weak magnitude of, and the limited individual variability in, the predictive relationship between perceived and assessed comprehension may be attributed to internal or external factors, the results of the present study indicate that individuals likely engage in fundamentally similar behaviours when providing judgements of comprehension. The observation of markedly similar behaviour across individuals, evidenced in both Study 1 and 2, allows for speculation concerning the conditions which may produce this, such as a reliance on a shared default source of information. It remains unclear what this default source of information may be. Nevertheless, these results contribute to a body of evidence which formal theoretical models of metacomprehension accuracy must be able to account for, with further research required in order to progress towards such a theoretical account. To explore whether internal or external factors, or a mix of both, may account for the weakly positive average association, evaluating the impact of reducing measurement error on the strength of the predictive relationship may be a productive approach for further research.

4.4.3 Limitations

It is important to note the limitations of the current study, relating to the experimental design and the measures of assessed comprehension and perceived comprehension. As discussed (see section 4.1.1), comprehending information within texts involves the use of comprehension-building processes, regardless of the position of the information with the semantic structure of the text. Consequently, comprehension outcomes relating to information which is more or less important within the macrostructure is expected to share variance. Indeed, this association was observed in the present study (see Figure 4.3). However, semantically central and peripheral comprehension outcomes were also associated with sufficient unique variance to permit the statistical model to be successfully fitted to the data. Nevertheless, the design used in Study 2 could be considered non-optimal, due to this association undermining the capacity to assess the relative contribution of semantically central and peripheral comprehension to individuals' judgements.

An alternative methodological approach, capable of resolving issue of shared variance in comprehension outcomes, is the error detection paradigm. Error detection tasks involve inserting an inconsistency to disrupt text comprehension (Baker, 1979; Markman, 1977, 1979). Similar to previous research (Baker, 1979; Baker & Anderson, 1982; Grabe et al., 1987; Yussen & Smith, 1990), comparisons between intact texts, texts with inconsistencies in semantically central information, and texts with inconsistencies in semantically peripheral information would permit the researcher to robustly evaluate whether the semantic centrality of the error, rather than the presence of an inconsistency alone, influences behaviour. However, researchers have previously discussed how the error detection task undermines Grice's (1975) 'Cooperative Principle', which states that readers' assume that text is coherent and relevant, resulting in the potential for unnatural reading behaviours (Ehrlich et al., 1999; Winograd & Johnston, 1980). As readers become aware of text inconsistencies, their purpose for reading may develop into overt coherence verification. Given the importance of reading goals in text comprehension processes (Kintsch & van Dijk, 1978; van den Broek et al., 2011), this may influence both the macrostructure derived from the text and the formation of metacomprehension judgements.

In addition to the potential lack of generalisability to judgements made on intact texts, introduced through using an error detection task, this approach requires the identification of semantically central and peripheral information which may be challenging. As discussed in the context of theoretical implications, it is arguably unclear how effectively the comprehension questions used in the present study assessed information which an average reader would be highly likely to consider semantically central and peripheral. Furthermore, regardless of the position of the information within the semantic structure of the text, in any metacomprehension study, the comprehension questions used to measure understanding do not exhaustively assess the information contained within the text. The measure of an individual's understanding of the text will, therefore, be inaccurate to some extent, and may seriously undermine our ability to investigate the basis of metacomprehension judgements (Weaver, 1990; Wiley et al., 2005). Given that constructing and testing near exhaustive sets of comprehension questions across multiple texts is not feasible, it may be more pragmatic for researchers to attempt to reduce measurement error by narrowing the scope of the metacomprehension judgement, from the level of the whole text to some smaller unit of meaning (e.g., Dunlosky et al., 2005; Dunlosky et al., 2007).

More generally, comprehension questions themselves may undermine the capacity to estimate the predictive accuracy of metacomprehension judgements. In the present study, the estimated question-level intercept variability indicated that there were clear differences in difficulty between comprehension questions. This variability may arise from differences in the difficulty of comprehending text information, variously due to grammatical and syntactical complexity, vocabulary, required background knowledge, or the need to draw complex inferences (Graesser, et al., 1994; Kintsch & van Dijk, 1978). However, attributes of question construction and response processes may also contribute to this, such as the comprehensibility of the question stem and response options, the method of response, or the plausibility of distractors (Ebel & Frisbie, 1991; Graesser et al., 2009; Oakhill et al., 2014; Ozuru et al., 2013). The measure of assessed comprehension does not separate these sources of variance: comprehension scores are a composite measure of text comprehension, question comprehension and response performance processes (Collins et al., 2020; Eason et al., 2012; Nation & Snowling, 1997). If valid information is not available to individuals concerning the nature of the upcoming comprehension test, judgements of comprehension will be unlikely to accurately incorporate an evaluation of the impact the measurement process on comprehension performance, undermining the predictive association between these measures (Griffin, Wiley & Thiede, 2019; Thiede et al., 2011; Wiley et al., 2005). Controlling or quantifying this source of measurement error, without influencing the process of forming metacomprehension judgements, remains a challenge within this area of research.

4.4.4 Conclusion

Study 2 found a weakly positive predictive relationship between perceived comprehension and assessed comprehension, with this relationship not found to reliably depend on the importance of the assessed information within the semantic structure of the text. Limited individual variability in this relationship was observed and differences in background knowledge, reading ability and reader-based standards of coherence did not reliably influence the relationship between perceived and assessed comprehension. Judgements of perceived comprehension, therefore, provide some information of predictive validity which is informative of both an individual's understanding of semantically central information and peripheral details. As such, these judgements may be useful in improving the likely comprehensibility of health-related information. Given the limited variability in the predictiveness of judgements on comprehension outcomes across individuals, it is likely that the judgements of individuals on reader panels are similarly predictive of comprehension.

To the extent that the estimated relationships may be informative of processes occurring during the formation of metacognitive judgements, it would appear that the ability to construct a coherent macropropositional structure of the text does not specifically influence judgements. The observed results are not fully accommodated by existing accounts of metacomprehension judgements, and instead suggest that individuals similarly utilise a primary source of information to evaluate comprehension quality with limited differences in the judgement process.

Further research is required to address the highlighted issues. Specifically, identifying approaches to reduce error arising from the measurement of perceived and assessed comprehension. This would permit researchers to determine whether the weakly positive predictive relationship is due to factors which internally underlie the judgement process or attributes of the response and measurement process. Importantly, this may yield implementable methods which increase the informativeness of metacomprehension judgements in the evaluation of health information. This issue is considered in the following study in the present research.

5. Study 3

In this chapter, the motivation for the third study (Study 3) is discussed, leading to the identification of one research question to be addressed. In the method section, firstly, the identification of the required sample size is considered in a design analysis. Following this, a pilot study and the participants, materials, and procedure of Study 3 are described. The presentation of results is separated into a preliminary data inspection, the main planned analysis, and an analysis of the sensitivity of the main findings to alternative analytical choices. Lastly, the findings of Study 3 are discussed.

5.1 Introduction

Consistent with previous research (Dunlosky & Lipko, 2007; Lin & Zabrucky, 1998; Maki, 1998; Prinz et al., 2020a; Yang et al.,2022), in both Study 1 and 2, a weakly positive predictive relationship between perceived and assessed comprehension was observed. The estimated magnitude of the association, therefore, appears relatively robust. This indicates that metacomprehension judgements typically provide some insight into comprehension outcomes on health-related texts, offering some capacity to discriminating between texts which are somewhat more or less well understood. However, in contrast to the considerable individual-level variability observed in previous research (Chiang et al., 2010; Glenberg & Epstein, 1985; Jee et al., 2006), limited differences in the strength of the predictive relationship between individuals were found in both Study 1 and 2. This discrepancy is likely due to the use of two-step analytic approaches adopted within previous metacomprehension research which, given the quantity of observations collected at the participant-level, exaggerate individual-level estimates (Schönbrodt & Perugini, 2013; Yarkoni, 2009). Based on analyses which use partial pooling to efficiently estimate participant-level variance, Study 1 and 2 indicate that the vast majority of individuals can be expected to provide judgements which are weakly predictive of their comprehension, on average.

In accounting for the weak association between measures of perceived and assessed comprehension, theories of metacomprehension typically emphasise inaccuracies related to the internal judgement process. These accounts identify a lack of available cues or the failure to use of valid cues in forming judgements, such that the internal evaluation is poorly predictive of comprehension outcomes (Griffin et al., 2009; Wiley et al., 2016). For example, inaccurate judgements may result from failing to monitor one's ongoing reading process at an appropriate level (Dunlosky, Rawson & Hacker, 2002). Further, while cues may relate to the text, they may not provide valid information. Individuals may form inaccurate judgements based on sources of information such as the perceived novelty of the text, expectations of a memory-based test, or the quantity of information which can be recalled (Griffin, Wiley & Thiede, 2019; Koriat, 1995; Thiede et al., 2010). Alternatively, multiple cues may be combined to yield an inaccurate judgement through the use of an anchoring-adjustment heuristic, in which experience-based cues do not sufficiently adjust the judgement or inaccurate belief-based cues dominate the evaluation (Linderholm et al. 2008; Zhao & Linderholm 2008). In either failing to effectively utilise or to generate relevant cues, accounts of metacomprehension generally suggest that low metacomprehension accuracy occurs due to the internal evaluation of comprehension being based on sources of information which lack correspondence with the underlying state of text comprehension.

Less frequently, researchers have discussed the importance of factors that are external to the comprehension evaluation process and that may account for the weakly positive alignment between perceived and assessed comprehension. While the internal process of comprehension evaluation may, in fact, produce an accurate appraisal, low metacomprehension accuracy may be observed due to error introduced in the measurement process. Previously, the vulnerability of the measurement of text comprehension to measurement error has been highlighted (Weaver, 1990; Wiley et al., 2005). To avoid capturing only prior knowledge or surface memory of the text, researchers have argued that a valid measure of understanding must use questions which are inference-based and pertain to the situation-model level of the text representation (Griffin, Wiley & Thiede, 2019; Wiley et al., 2005; 2016). Further, it has been argued that a valid measure of comprehension is underpinned by an adequate number of comprehension questions which sufficiently capture an individual's breadth of understanding of a text (Weaver, 1990; Wiley et al., 2005). Given that a small number of questions can target only a limited portion of a text, the largest possible discrepancy between judgement and performance is most likely to occur when comprehension, measured using a single question, is evaluated against a judgement of comprehension concerning a text (Lin & Zabrucky, 1998; Weaver, 1990; Wiley et al., 2005).

In obtaining a valid measure of text comprehension, it could be argued that it is not the number of comprehension questions used to assess comprehension which undermines metacomprehension accuracy, but the appropriateness of a certain number of questions given the scope of the judgement prompt. For example, if a metacomprehension judgement concerned a single piece of information within a text, it may be considered reasonable to use a single question to assess whether a reader's understanding of the information had reached some threshold. In contrast, using a single question to assess comprehension relative to a judgement prompt targeting the whole text would produce a considerable mismatch in the 'grain size' of the judgement and performance measures, thereby undermining the association between perceived and assessed comprehension (Dunlosky et al., 2005; Weaver, 1990; Wiley et al., 2005).

While increasing the number of comprehension questions to provide a sufficiently exhaustive measure of overall text comprehension presents a number of challenges, in both constructing the set of comprehension questions and collecting a suitably sized sample of responses to such questions (Griffin, Mielicki & Wiley, 2019), decreasing the scope of the metacomprehension judgement is readily achievable. By narrowing the focus of the metacomprehension judgement, the correspondence between the scope of the judgement and the coverage of the assessment of comprehension should be increased, thereby improving metacomprehension accuracy (Weaver, 1990; Wiley et al., 2005). In evaluating this suggestion, the limited research that has considered the relationship between the scope of the judgement and metacomprehension accuracy is discussed in the next section.

5.1.1 Scope of the Judgement

Within studies which have examined the relationship between the scope of metacomprehension judgements and task performance, metacognitive judgements are loosely grouped into two categories. Judgements which relate to multiple pieces of information are typically referred to as 'global judgements' (Dunlosky & Lipko, 2007), while judgements which relate to individual items or a subset of a body of stimuli are more variably referred to as 'local judgements', or item- or term-specific judgements (Dunlosky et al., 2005; Nietfeld et al., 2005). These categories are not well-defined, however, with the quantity of information covered within each judgement type varying between studies. The defining feature between these two judgement types, across studies, is that global judgements refer to prompts which target a greater quantity of text than local judgement prompts.

As discussed, researchers have previously argued that a lack of correspondence between the scope of the judgement and the coverage of the performance measure negatively impacts on metacomprehension accuracy (Lin & Zabrucky, 1998; Weaver, 1990; Wiley et al., 2005). However, independent of the coverage of the comprehension questions, the quantity of information targeted in a judgement prompt may also influence the internal processes of comprehension evaluation. It has been suggested that individuals may evaluate a sample of information when forming a global judgement of comprehension (Dunlosky et al., 2005). In sampling from available cues, whether related directly to the text or subjective experiences, error is introduced into the evaluation process, reducing the accuracy of the evaluation (Dunlosky et al., 2005; Händel & Dresel, 2018; Koriat, 1995; Lefèvre and Lories, 2004). The difficulty of providing a global evaluation has similarly been suggested to reduce judgement accuracy, by prompting individuals to rely on cues which are not directly related to the reading experience (Dunlosky et al., 2005; Dunlosky & Lipko, 2007; Händel et al., 2020).

To date, empirical studies which have examined the relationship between the scope of the judgement and metacomprehension accuracy have provided mixed results. Evaluating the effects on judgement magnitudes and response accuracy, research has found that either global judgements are more accurate than local judgements (Nietfield et al., 2005), or that there is little evidence of a difference (Dunlosky et al., 2005; Schraw, 1994). In contrast, predictions of paired-associate recall made at the level of each item have been found to be more accurate than a single prediction across all items, particularly when predictions are delayed (Connor et al., 1997). Similarly, conditional on a pre-judgement retrieval attempt, term-specific metacomprehension judgements have been found to produce a higher association between perceived and assessed comprehension than judgements made at the level of the whole text (Dunlosky, Rawson & McDonald, 2002; Dunlosky et al., 2005). However, research by Lefèvre and Lories (2004) failed to find such a difference in the metacomprehension accuracy of judgements made at the text-level and the paragraph-level, while Vössing and Stamov-Roßnagel (2016) found a more complex pattern of metacomprehension accuracy for chapter-level and specific-concept judgements, dependent on first eliciting a judgement of perceived difficulty.

Given the small number of studies in this research area that have yielded mixed results, the relationship between judgement scope and metacomprehension accuracy remains unclear. Further the extent to which these findings can inform the likely predictive accuracy of reader panellists' judgements is further limited by the experimental design. Eliciting predictions of performance, rather than judgements of understanding, may influence the judgement process by making expectations about an upcoming test salient (Griffin, Wiley & Thiede, 2019; Linderholm & Wilde, 2010). In addition, paragraphs contain multiple pieces of information that are likely not exhaustively assessed in the comprehension questions. Research that elicits judgements at a paragraph-level, therefore, may not effectively evaluate the difference between global and local judgements. Furthermore, as discussed in Chapter 2 (section 2.2), these studies use non-optimal analyses which involve an improper treatment of uncertainty and risk bias in estimation.

Considering whether reducing the scope of the judgement influences the predictive relationship between perceived and assessed comprehension would be highly informative of the utility of the reader panel review process. Should judgements made of specific pieces of information be more informative of understanding than judgements made of whole texts, health practitioners would be provided with a simple and implementable method to improve the text evaluation process. Further, this would provide an alternative approach which permits exploration of whether key details within health-related texts, which may be crucial for positive health-outcomes, are likely to be understood. In addition, we can evaluate the extent to which measurement error may be introduced, due to a mismatch between the judgement scope and coverage of the comprehension questions, by considering potential differences in predictive accuracy of local and global metacomprehension judgements.

5.1.2 Research Aims of Study 3

To explore whether the weak, on average, alignment between perceived and assessed comprehension observed in Study 1 and 2 may be attributed to a lack of alignment between the scope of the judgement and the performance measure, Study 3 was designed to consider whether the predictive relationship between perceived and assessed comprehension may be influenced by the scope of the metacomprehension judgement. Specifically, Study 3 sought to compare the association between metacomprehension judgements and assessed comprehension when judgement prompts elicit global judgements, made at the level of the whole text, or local judgements, made at the level of individual idea units within texts (Dunlosky & Lipko, 2007). Study 3, therefore, aimed to address one research question:

> **RQ5:** Are judgements of comprehension made at the of the whole text, or at the level of an idea within a text, better predictors of the comprehension of information on health-related texts?

Addressing the above research question requires a comparison of the predictive association, given global and local metacomprehension judgements, between perceived and assessed comprehension. Two methods of statistical comparison were chosen to robustly evaluate evidence for a difference, motivated by the potential influence of shared variance between global and local judgements. The analysis will comprise i) a comparison of the unique variance attributable to each of the two judgement types and ii) a comparison of whether either judgement type in isolation may be preferred. Firstly, in a model comprising both global and local perceived comprehension ratings ('joint model'), the magnitudes of the slope coefficients of these variables will be compared. Non-overlapping 80% HDIs for the effects of interest will be considered as indicating a reliable difference in the magnitude of the effects (Makowski et al., 2019). Secondly, in two separate models, where either global judgements ('global model') or local judgements ('local model') are specified as the predictor of comprehension, the relative predictive performance of these two models will be compared. In this analysis, an estimated difference in the ELPD of the two models of greater than four, with non-overlapping intervals constructed from the ELPD difference plus or minus three times the standard error, will be considered as indicating a reliable difference in the predictive performance of the models (Sivula et al., 2020).

5.2 Method

Ethical approval for this study was gained (for piloting and data collection) in July 2021. Piloting and data collection commenced in October 2021. A detailed preregistration for this study, including materials and analysis plan, was uploaded to an OSF repository in October 2021 (https://osf.io/asg6b/).

5.2.1 Design Analysis

To evaluate the capacity to estimate the effect of interest with accuracy and precision, as described in Chapter 2 (section 2.3.1), an analysis of the proposed design for Study 3 was conducted under varying sample sizes. The analysis for Study 3 will examine whether a difference is observed in i) the estimate magnitudes within the joint model or ii) the predictive performance of the global model and local model. Given potential dependence between global judgements and local judgements, a considerable number of assumptions would be required, with limited supporting evidence for justification, to appropriately conduct the simulations. Due to this uncertainty, it was considered that using simulation to examine the probability of observing evidence for either difference was not maximally informative. Instead, the design was evaluated by considering the accuracy and precision of the estimated effect for global and local judgements given separate models. The sample size of Study 3 is considered adequate when the probability of achieving the target level of accuracy and precision in estimation is expected to occur in at least 80% of hypothetical studies.

The research question, RQ5, in Study 3, given separate judgement models, concerns two parameters of interest: i) the effect of global perceived comprehension and ii) the effect of local perceived comprehension. To determine the sample size required to estimate these effects with accuracy and precision, a multilevel logistic regression model was used to simulate response data. In addition to this, contrasting with the simulations conducted in

Study 1 and 2, multilevel cumulative probit models were used to simulate judgements of perceived comprehension. This alternative approach to simulating judgements was afforded by the data obtained in Study 2 and the anticipated similarity in the experimental materials between Study 2 and 3. Selection of model parameter values were informed by Study 2 and previous research, were appropriate. Further details of the simulation and model specification are provided below.

Perceived Comprehension Simulation Models. To inform the simulation of judgements of perceived comprehension, two classes of regression models were fitted to Study 2 data: i) a multilevel cumulative probit regression model and ii) a multilevel sequential ratio regression model. As the cumulative model provided a slightly better fit with lower model complexity, this model was selected for the simulation of global judgements of perceived comprehension. In the absence of data to inform the simulation of local judgements of perceived comprehension, the cumulative probit model was also used as the basis for simulating observations of local judgements. A full description of the global and local judgement simulation models is provided in Appendix M.

Assessed Comprehension Simulation Models. Two simulation models were specified to generate observations of assessed comprehension, comprising of either simulated global judgements or local judgements as the predictor of response accuracy.

Global Comprehension Simulation Model. In the global comprehension simulations, the binary response Y denotes whether or not an individual i, reading text j, answered a comprehension question k correctly ($Y_{ijk} = 1$) or incorrectly ($Y_{ijk} = 0$). Letting P_{ijk} be the probability of observing a correct response $P\left(Y_{ijk} = 1\right)$, this event can be expressed at the result of a Bernoulli trial:

$$Y_{ijk} \sim \text{Bernoulli}\left(P_{ijk}\right),$$

where i = 1, …, I; j = 1, …, 13 and k = 1, …, 6. Assuming a logit-link function, the explanatory variables of interest and hierarchical sources of variance were linked to the response:

$$\log\left(\frac{P_{ijk}}{1-P_{ijk}}\right) = \beta_0 + (\beta_1 + u_{1i})GlobalJudgement_{ij} + u_{0i} + v_j + w_k, \qquad (11)$$

where $GlobalJudgement_{ij}$ refers to the observed global rating of perceived comprehension from individual i in response to text j, with values corresponding to the seven available rating responses: $GlobalJudgement_{ij} = (0,1,2,3,4,5,6)$; β_0 refers to the intercept (baseline log odds of observing success); β_1 refers to the population-level effect of global perceived comprehension; u_{1i} refers to individual-level variability in the effect of global perceived comprehension; and u_{0i}, v_j, w_k and refer to intercept variability at the level of the individual, text and question, respectively.

Model-fitted estimates of parameters, obtained using the data collected in Study 2, were used to inform parameter values. The baseline probability of answering a comprehension question correctly, given the lowest rating of global perceived comprehension, selected on this basis, was $\beta_0 = 0.27$. The population-level effect of global perceived comprehension judgements was $\beta_1 = 0.06$.

As it was anticipated that the same text and question stimuli would be used in Study 3, model-fitted estimates from Study 2 for the text-level and question-level intercept deviates were used for v_j and w_k. In contrast, individual-level variability in the intercept u_{0i} and effect of global perceived comprehension u_{1i} were simulated. Due to high uncertainty in the covariance parameter in Study 2, these deviates were simulated from univariate normal distributions based on model-fitted estimates:

$$u_{0i} \sim \mathrm{Normal}(0, 0.7),$$

$$u_{1i} \sim \mathrm{Normal}(0, 0.1).$$

Local Comprehension Simulation Model. In the local comprehension simulations, the binary response Y denotes whether or not an individual i, reading text j, answered a comprehension question k correctly ($Y_{ijk} = 1$) or incorrectly ($Y_{ijk} = 0$). The corresponding logistic model can be expressed as:

$$Y_{ijk} \sim \text{Bernoulli}\left(P_{ijk}\right),$$

$$\log\left(\frac{P_{ijk}}{1 - P_{ijk}}\right) = \beta_0 + (\beta_2 + u_{2i})LocalJudgement_{ijk} + u_{0i} + v_j + w_k, \qquad (12)$$

where $LocalJudgement_{ijk}$ refers to the local rating of perceived comprehension from individual i in response to idea k from text j, with values corresponding to the seven available rating responses: $LocalJudgement_{ijk} = (0,1,2,3,4,5,6)$. For clarity between models, β_2 and u_{2i} refer to the population-level effect of local perceived comprehension and individual-level variability in this effect, respectively. The meaning of all other parameters is as defined for (11).

The selection of parameter values and simulation of group-level deviates was as described in the global comprehension simulation model (based on model-fitted estimates from Study 2) for all but one parameter: β_2. A conservative magnitude was selected to express the difference in the effect of perceived comprehension between global and local judgements. An increase of +0.1 in the overall change in probability was selected for local judgements compared to global judgements, from the minimum to maximum metacomprehension judgement. Expressed as a change in the log odds per unit increase in the local judgement of perceived comprehension, $\beta_2 = 0.14$.

Individual-level variability in the effect of local perceived comprehension u_{2i} was simulated using a univariate normal distribution, with an equivalent standard deviation to variability in the effect of global perceived comprehension:

$$u_{2i} \sim \text{Normal}(0, 0.1).$$

Simulation Procedure. The simulation was conducted using the HEC facility at Lancaster University using R (R Core Team, 2019). The simulation was conducted under varying numbers of participants, with six values for the total number of participants considered: I = (100, 125, 150, 175, 200, 225). The number of texts and questions per text were fixed to 13 and 6, respectively. A dataset of observations for each sample size was simulated 1000 times.

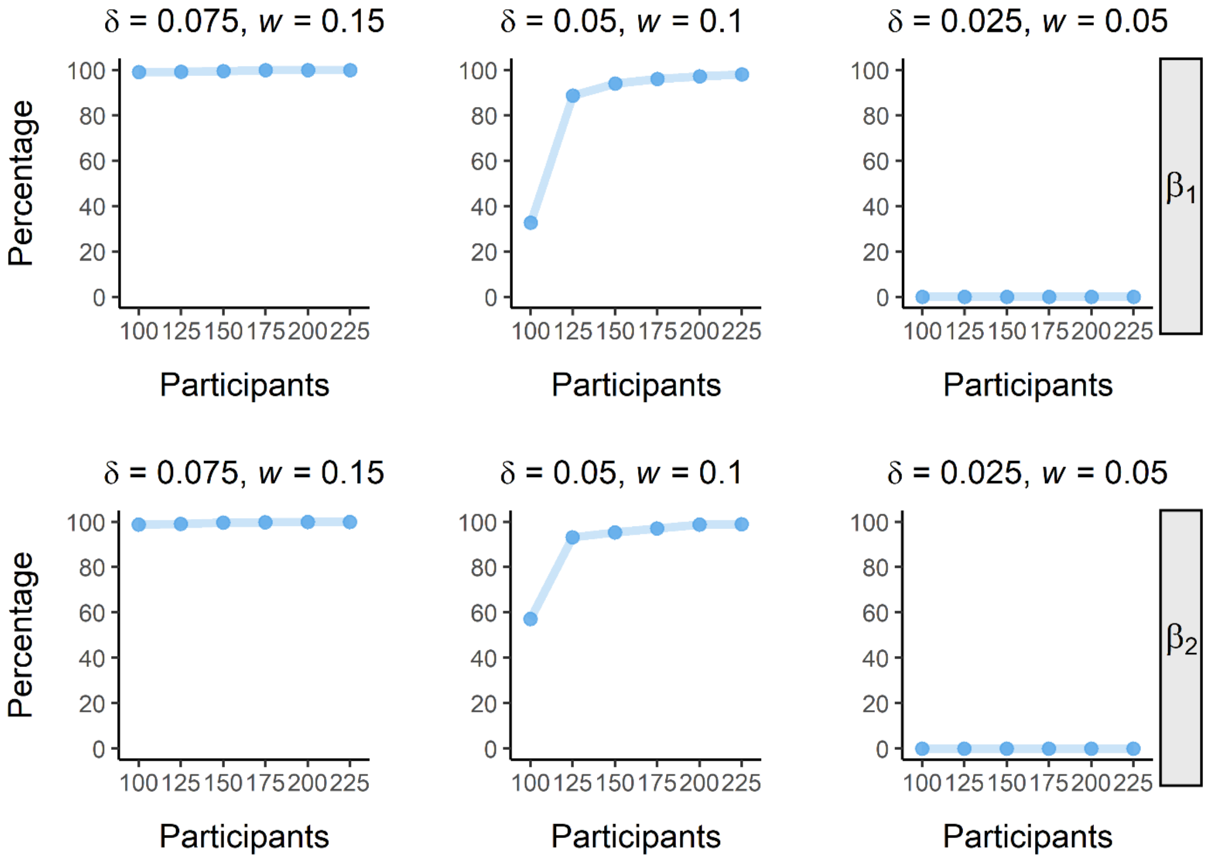
Model Fitting. Multilevel logistic models, featuring either global (11) or local (12) judgements of as the predictor, were fitted to each simulated dataset to estimate the effects of interest. Models were fitted using a Bayesian estimation framework, via the brms package (Bürkner, 2017, 2019) and Stan (Carpenter et al., 2017). Weakly informative priors were specified for the model parameters. For the population-level effects (β_0, β_1, β_2), normal distributions with mean 0 and standard deviation 10 were specified. Half-student-t distributions with the values of the degrees of freedom, location and scale parameters set to 3, 0, and 10 for the group-level effects (u_{0i}, u_{1i}, u_{2i}, v_j, w_k). An LJK correlation distribution with shape parameter of 1 was used at the prior on the covariance between participant-level variance in the intercept and the effects of global and local perceived comprehension. From each resulting model fit, estimates for the mean of the posterior distribution and the 90% credible intervals for the effects of interest were obtained.

Simulation Results. The capacity to estimate the effects of interest, for a given number of participants, was operationalised in terms of the probability of estimating β_1 and β_2 with accuracy and precision. Due to the smaller magnitude of effects specified in the design analysis for Study 3, different pairs of values for δ and w were selected to represent **lower** ($\delta = 0.075, w = 0.15$), **middle** ($\delta = 0.05, w = 0.1$) and **higher** ($\delta = 0.025, w = 0.05$) levels of accuracy and precision in estimation. For the **lower** level of accuracy and precision, for both global and local judgements, accepted values for the point estimate include negative values and values more than double and less than half of the true effect. For the **middle** level of accuracy and precision, for global judgements, accepted values for the point estimate exclude negative values and values more than double the true effect. For the **middle** level of accuracy and precision, for local judgements, accepted values for the point estimate exclude values less than half the true effect, and, separately, negative values and values more than double the true effect are excluded from the 90% credible interval. For the **higher** level of accuracy and precision, for both global and local judgements, accepted values for the point estimate which are more than double or less than half are excluded and, separately, negative values and values more than half of the true effect are excluded from the 90% credible intervals. The probability of successfully achieving accuracy and precision in estimation, across varying δ and w, given each total number of participants simulated, is shown in Figure 5.1.

From left to right in Figure 5.1, decreasing values of δ and w correspond to higher demands on accuracy and precision in estimation for β_1 (top row) and β_2 (bottom row). The simulation indicated that for the **lower** level for accuracy and precision, all sample sizes considered were sufficient to successfully achieve the goal of estimation on more than 80% of occasions. In contrast, none of the sample sizes considered were found to meet the goal of estimation given the **higher** level of accuracy and precision. For the **middle** level of accuracy and precision, sample sizes greater than 100 participants were required to achieve the estimation goal on 80% of occasions. In addition, greater numbers of participants were required to achieve the **middle** target level of accuracy for β_1 compared to β_2.

Figure 5.1

Probability of Achieving Varying Levels of Accuracy and Precision in Estimation in Study 3 Simulation



Note. The lines show the percentage of simulations achieving the specified target level of accuracy (δ) and precision (w) under varying numbers of participants (horizontal axis). From left to right, the level of accuracy and precision in estimation increases. The upper and lower rows of the plot correspond to β_1 and β_2, the population-level effect of perceived comprehension for global judgements and the population-level effect of perceived comprehension for local judgements, respectively.

Using the same sample sizes for texts and questions per text as Study 2 (13 texts and six questions per text), a sample size of 125 participants was estimated to provide a greater than 80% probability of achieving the **lower** and **middle** level of accuracy and precision considered for both β_1 and β_2. This suggests that 125 participants would likely provide adequate accuracy and precision in estimating the effects of both global and local perceived comprehension ratings. However, this simulation evaluates the probability of estimating individual effects within given models with accuracy and precision, whereas the research question for Study 3 will be address using comparisons of the magnitudes of effects within, and the relative predictive performance between, models.

Given the divergence in the context for which the simulation is maximally informative and the planned analyses, uncertainty is introduced with respect to how informative the results of the simulation are of the outcomes of the planned analysis. To help mitigate potential limitations in addressing the research question, due to insufficient volumes of data, a larger sample size was selected. Given limited additional information to inform this choice, whilst balancing the concerns identified and available funds to conduct the research, an additional 45 participants above the sample size suggested by the simulation was selected ($I = 170$).

5.2.2 Pilot

Given the similarity with the design of Study 2, a 'within data collection' pilot was conducted to evaluate the proposed study design. This consisted of pausing data collection after the first 25 participants had completed the study and checking responses to detect any major problems with the experimental apparatus and to confirm that the duration of the experimental session was as anticipated. The first 25 participants were aged 21 to 50 (*M* = 31.16, *SD* = 7.93), with 22 identifying as female and three as male. No formal analyses were run on this subset of the sample. As no concerns were identified, data collection was resumed until the target number of participants had been reached. Since the recruitment process, materials and procedure were identical in the pilot and Study 3, these will be discussed in the context of the main data collection below.

5.2.3 Participants

A sample of 171 participants was recruited using the online platform Prolific (due to a technical error, one additional participant above the target sample size of 170 was recruited). Participation was limited to UK nationals that had not taken part in any tasks in Study 1 or 2. Recruitment was conducted over several days to capture a greater spread of participants on the Prolific system. Consistent with Study 1 (see Chapter 3; section 3.2.3), exclusion criteria specified that participants would be excluded and replaced if they demonstrated limited evidence of engagement with the task. Reading times were used as a proxy measure of engagement, with insufficient engagement defined as reading times less than the minimum duration given a reading rate of 600 wpm. The impact of participant exclusion on the results are explored in the sensitivity analysis (section 5.3.3).

Following the sampling procedure, participants were recruited until 170 submissions which met the acceptance criteria were obtained (page submission times on all 13 health-related texts corresponding to a reading rate greater than 600 wpm). In total, 197 participants were recruited. Responses from 26 participants who met the exclusion criteria were removed and replaced. Accepted participants (*I* = 171) were aged from 21 to 77 (*M* = 44.15, *SD* = 13.24), with 90 identifying as female and 81 as male. All participants who fully completed the study received a payment of £5.00.

5.2.4 Materials

Health Texts. The same health-related texts developed in Study 2 were used in Study 3 (see Appendix I). Descriptive information for the texts is provided in Table 4.1 (Chapter 4; section 4.2.4).

Comprehension Questions. The comprehension questions used in Study 3 were based on those developed in Study 2. In Study 2, comprehension questions were designed to specifically target information considered likely to be semantically central or semantically peripheral in the reader’s macrostructure of the text. However, in Study 3, question types were grouped together to provide a measure of assessed comprehension consisting of six questions per text. This was considered reasonable given similar observed levels of performance accuracy on semantically central and peripheral comprehension questions in Study 2 and that the centrality of the information was not of relevance in Study 3.

The set of questions developed in Study 2 was reviewed prior to use in Study 3, using two methods. Firstly, the 78 questions were re-evaluated by three researchers to confirm that the comprehension questions and correct responses tested understanding of information contained in the text and that the distractors could not be considered correct or partially correct based on the text. Secondly, using the data collected in Study 2 which met the acceptance criteria (*I* = 225), response option selection rates were evaluated for indications of problematic questions. This involved identifying questions which the top performing quartile of participants answered incorrectly on 50% or more occasions. Following this, response option selection rates were evaluated for indications of distractor bias. Any question which appeared to show bias in distractor selection was then examined further, by re-examining the information provided in the text and evaluating whether this bias was observed in the top performing 10% of participants. This procedure identified 11 questions which required alteration. As previously, modifications to questions and response options were separately reviewed by three experimenters and altered until all experimenters were satisfied. The set of questions with response options used in Study 2 are provided in Appendix K, with the questions which were revised in Study 3 provided in Appendix N.

Perceived Comprehension. For each text in the metacomprehension task, one global judgement prompt and six local judgement prompts were designed to measure of perceived comprehension. To elicit global judgements of comprehension, the prompt 'Overall, how well do you understand the text?' was selected. This prompt was chosen for the text-level judgement as it showed a similar pattern of results in the sensitivity analysis as the prompt used in Study 1 and 2 ('How much of the text do you feel you understand?') but was considered more likely to encourage participants to consider comprehension of the text as a whole.

Prompts for local judgements targeted specific pieces of information within the text, each prefaced with the wording 'How well do you understand ...', followed by the target information. Prompts were constructed with caution, considering the potential influence they may have on participants' response processes. To reduce the potential for a correct response to be generated by recall of the prompt or reduced plausibility of response options, prompts were worded to avoid stating the correct answer to the subsequent comprehension question or invalidate the distractors. Simultaneously, prompts were worded to limit ambiguity in the information offered for evaluation and to ensure that this information corresponded to the same information tested in the subsequent comprehension questions. Obtaining specificity in prompt wording while ensuring the comprehension question response options were not reinforced or invalidated was challenging to achieve for some prompts. In such cases, a prompt which offered the best compromise was selected (e.g., altered plausibility of some, but not all, response options, given the prompt wording). The judgement prompts were separately reviewed and altered until three experimenters were satisfied with the wording. The resulting set of 78 local prompts is provided in Appendix O.

For global and local judgements, ratings were captured using 7-point rating scale. The rating scale was presented as an unmarked line with a moveable slider, initially not shown on the scale until the participant interacted with the scale, to be dragged left or right to indicate judgement (the slider 'snapped' to the value closest to one of the seven ratings). The labels 'not at all well' and 'extremely well' were presented on the left and right of the scale, respectively.

5.2.5 Procedure

Participation in the study commenced when participants responded to the online invitation on Prolific.co. All tasks were computer-based and presented via Qualtrics. Participants provided consent and reported their age and gender prior to completing the tasks.

Participants first read each health text in the metacomprehension task and made judgements of comprehension immediately following the reading of each text. Texts were presented in one of four randomised orders. Immediately after reading a text, judgements of perceived comprehension were elicited in response to the six local judgement prompts, presented together on screen without the text. Following this, participants provided a single global rating of perceived comprehension, presented on a separate screen. Participants were invited to take a break if needed after reading and judging all the texts. Participants were then presented with the comprehension questions for each text, with the text present, in the same order that the texts were presented for reading. After completing the comprehension questions for all texts, participants were fully debriefed.

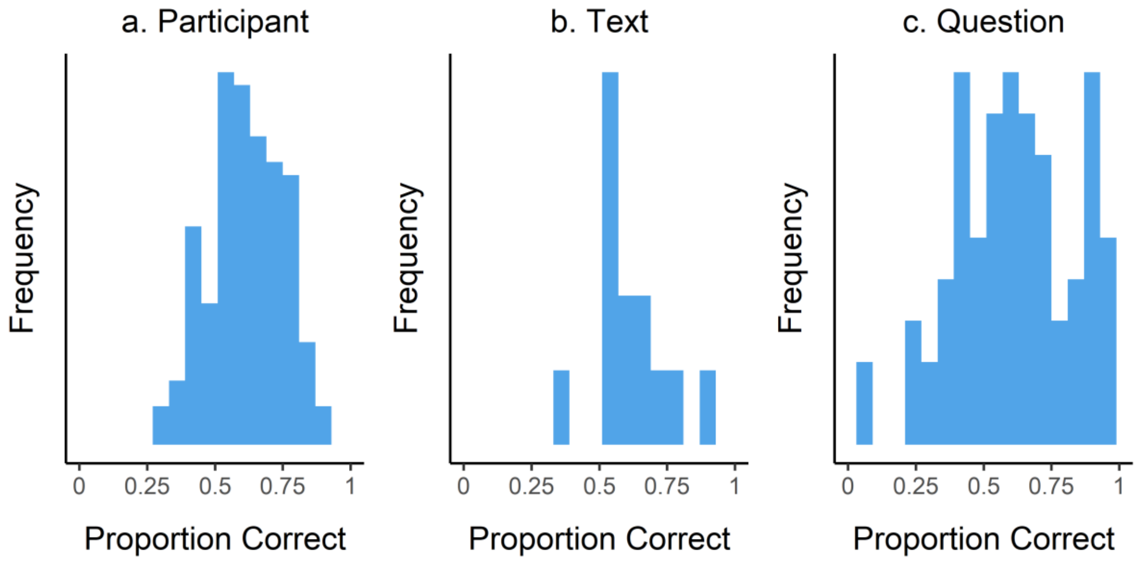
5.3 Results

5.3.1 Preliminary Data Inspection

Initial inspection of the responses which met the acceptance criteria indicated that performance on the multiple-choice questions was comparable to Study 2. Across participants, texts and questions, the proportion of comprehension questions answered correctly on the metacomprehension task was 0.61 (Study 2 average: 0.59). This corresponds to participants correctly answering between three to four of the total six questions per text, on average. Variability in comprehension question accuracy was evident between participants, texts and questions. Figure 5.2 displays the distributions of the proportion of questions correctly answered by participant (5.2a.), text (5.2b.), and question (5.2c.).

Figure 5.2

Proportion of Comprehension Questions Correctly Answered in the Metacomprehension Task in Study 3

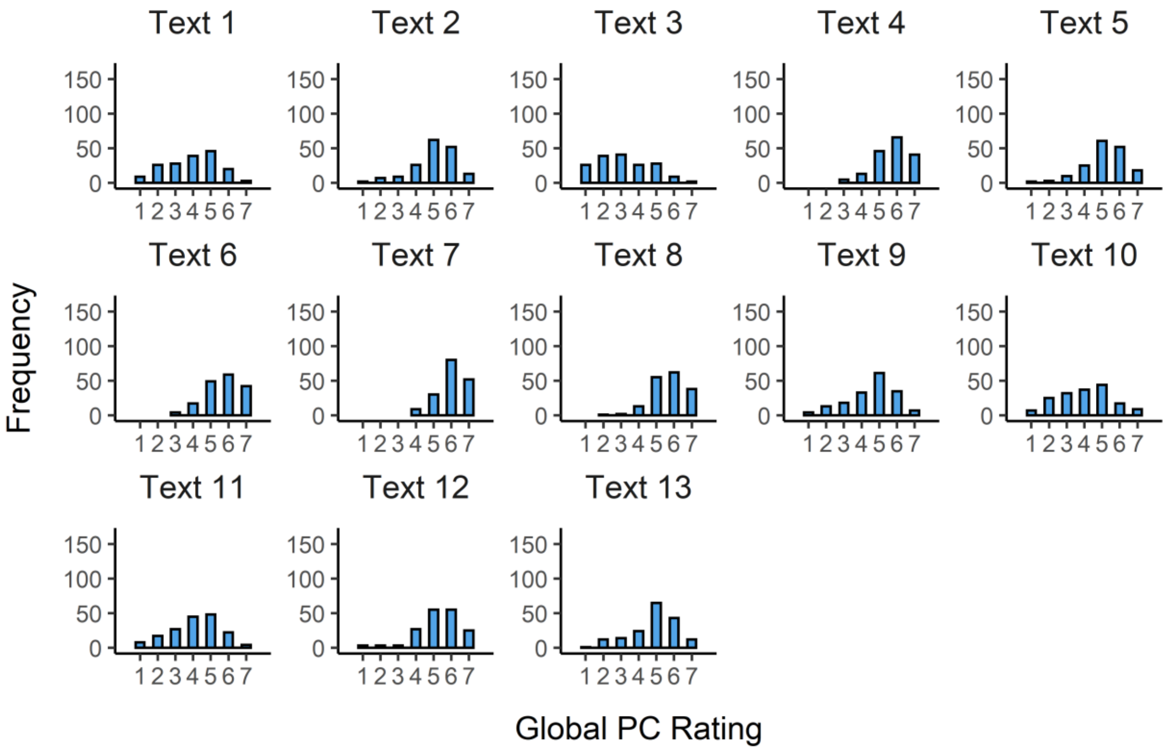


Note. Participant proportion correct (*I* = 171), text proportion correct (*J* = 13), and question proportion correct (*K* = 78).

Distributions of global perceived comprehension judgements, made at the level of a text, showed clear variability between texts, as shown in Figure 5.3. For the majority of texts, global judgements were negatively skewed, with participants reporting high levels of perceived understanding. For a smaller number of texts, global judgements appeared positively skewed or showed limited bias. Texts were typically judged as somewhat well understood, with the average global judgement being 4.86, across participants. Participants most frequently selected the 5th (29% of all global judgements) and 6th (26% of all global judgements) response options when providing judgements of overall understanding of a text. The extreme negative skew in the text-level metacomprehension judgements observed in Study 2 was, therefore, reduced in Study 3, with the highest rating of perceived comprehension selected almost three times less frequently in Study 3 compared to Study 2.

Figure 5.3

Distributions of Global Perceived Comprehension Judgements, by text, in Study 3

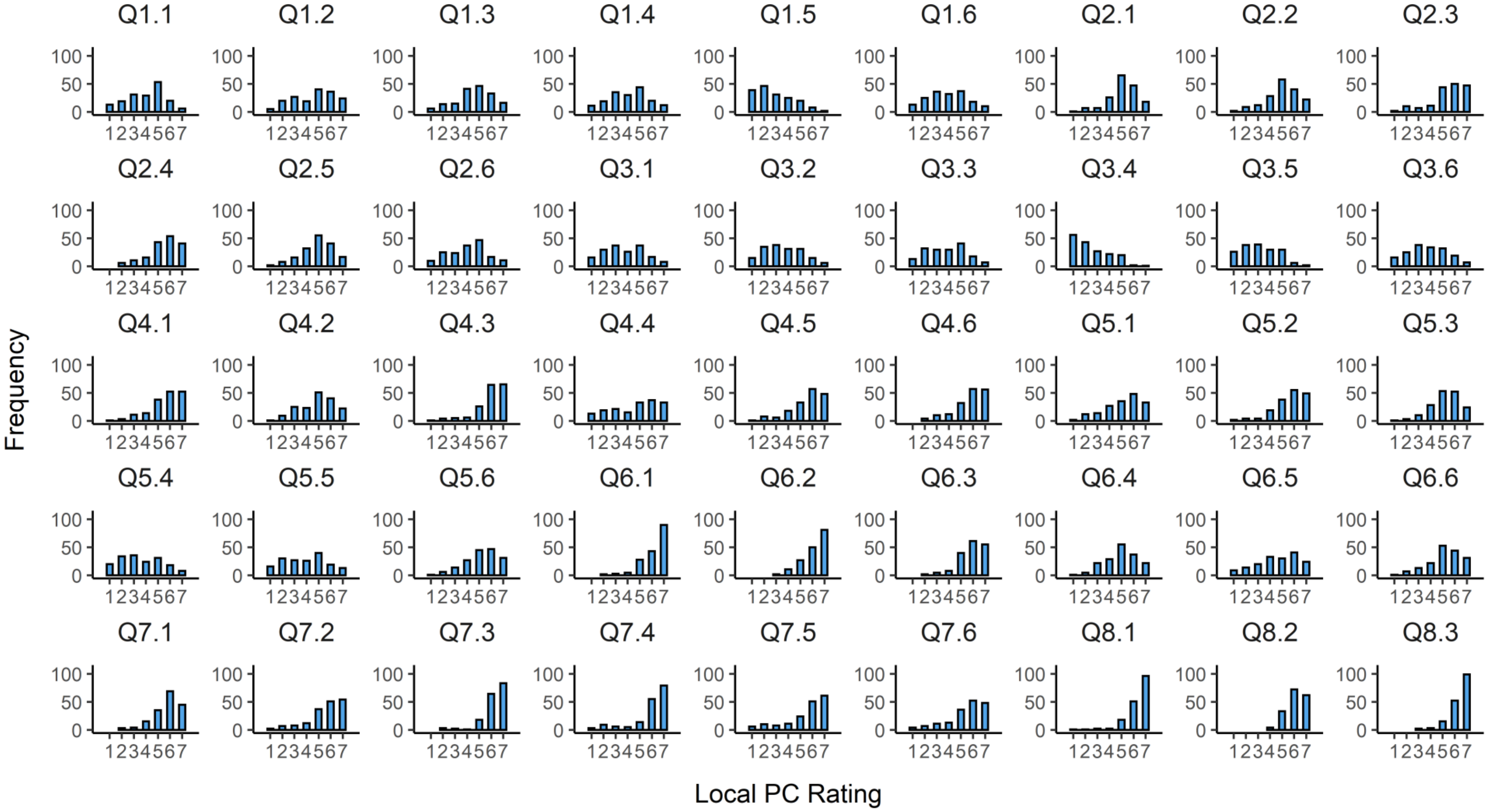


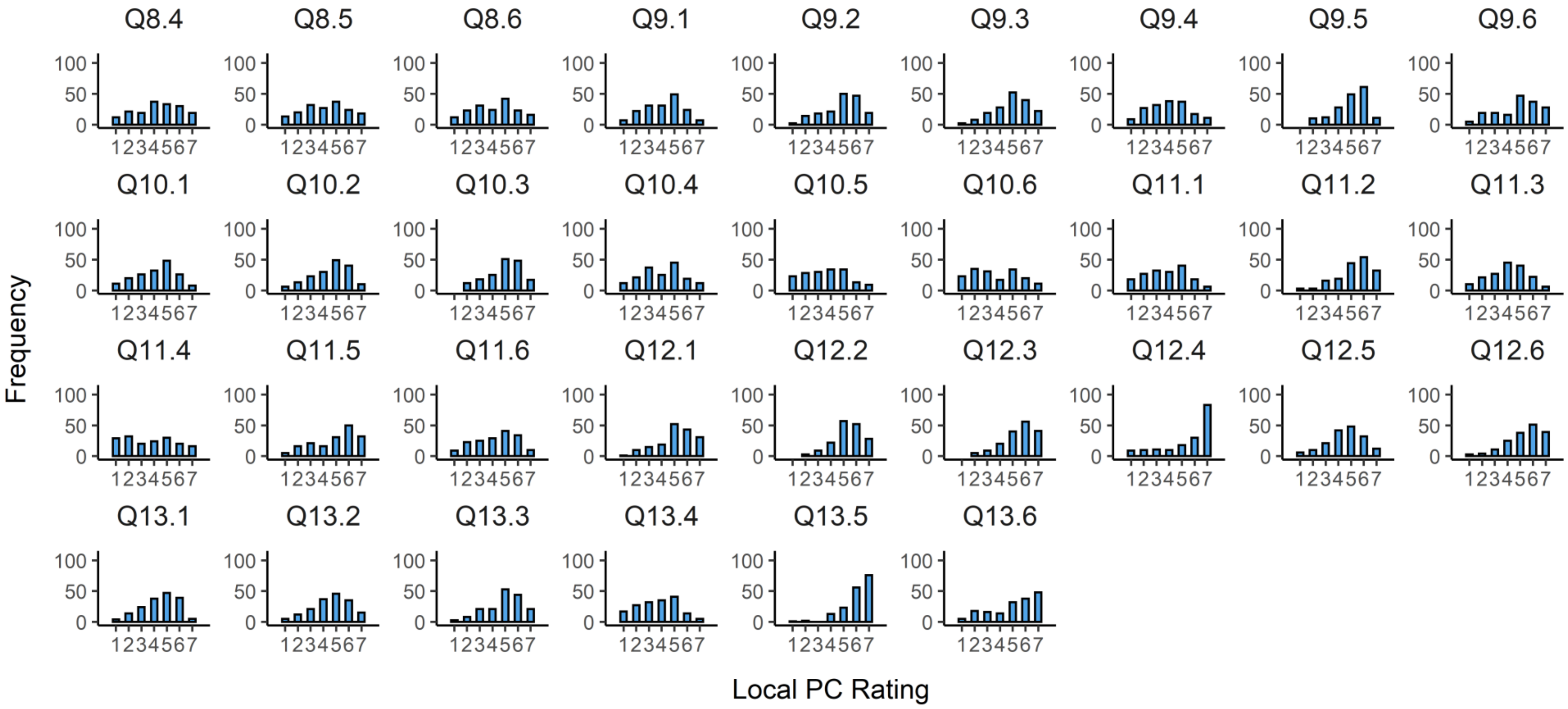
Note. PC = perceived comprehension.

Considerable variability was likewise observed in the distributions of local perceived comprehension judgements made at the level of specific information within a text. As each local judgement prompt corresponds to a question on the comprehension test, the distributions of local perceived comprehension judgements shown in Figure 5.4 are plotted according to the respective comprehension questions. The majority of local judgements distributions in Figure 5.4 show moderate to extreme negative skew, with participants predominantly reporting high levels of perceived understanding. For a number of local judgements, however, no clear bias in responses was observed across participants and in two cases a clear positive skew in responses was observed (labelled as Q1.5 and Q3.4 in Figure 5.4).

Figure 5.4

Distributions of Local Perceived Comprehension Judgements, by Corresponding Question, in Study 3





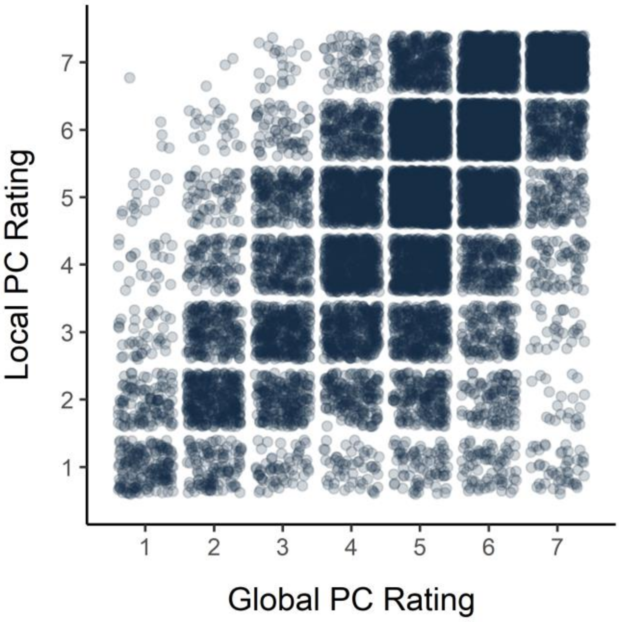
Note. ‘Q1.1’ refers to the first question on the first text. PC = perceived comprehension.

Overall, information within texts was typically judged as marginally less well understood than the texts overall, with the average local judgement being 4.78, across participants. Relative to distributions of global judgements, extreme ratings were observed slightly more frequently for local judgments. Across participants, the lowest and highest ratings were selected on 4% and 18% of occasions for local judgements, respectively, whereas this was 3% and 12% for global judgements.

Given the impact of shared variance on the estimation of coefficients in both the joint and separate models in the main analysis, the association between global and local judgements was also examined in the preliminary inspection of the data. Figure 5.5 shows a scatter plot of both judgement types, indicating a positive association between judgements of understanding at the level of the text and at the level of information within a text.

Figure 5.5

Scatterplot of Global and Local Perceived Comprehension Judgements in Study 3



Note. Points are plotted with jitter applied, adding small, randomised perturbations to the observations to increase the readability of the plot. PC = perceived comprehension.

5.3.2 Planned Analysis

To explore whether judgements of comprehension made at the level of information within a text, or at the level of the whole text, are better predictors of the comprehension of information on health-related texts, Bayesian multilevel logistic regression models were fitted in R (R Core Team, 2019) using the brms package (Bürkner, 2017, 2019) and Stan (Carpenter et al., 2017). The results of the analyses are discussed in the next section. Following this, a sensitivity analysis is reported to explore how various analytical choices may influence the findings.

RQ5. To address RQ5, two approaches were taken. Firstly, in a regression model with both global and local perceived comprehension judgements (joint model), the magnitudes of the estimated effects were compared. Secondly, in two models, where global and local perceived comprehension judgements were specified separately within each model (separate models), the relative predictive performance of the models was compared. Each of these analytic approaches are discussed in this section.

Joint Model. In the joint model, observations of assessed comprehension, from individual *i* on text *j* and question *k*, were used as the outcome variable Y_{ijk}. Judgements of perceived comprehension, elicited in response to both global and local judgement prompts, were standardised prior to model fitting and entered as predictors of the log odds of correctly responding to the comprehension question. Group-level variance in the intercept was estimated for participants u_{0i}, texts v_j and questions w_k. Participant-level variability in the effect of global u_{1i} and local u_{2i} perceived comprehension judgements was also included. The model fitted was therefore:

$$\log\left(\frac{P_{ijk}}{1-P_{ijk}}\right) = \beta_0 + (\beta_1 + u_{1i})GlobalJudgement_{ij} \qquad (13)$$

$$+ (\beta_2 + u_{2i})LocalJudgement_{ijk} + u_{0i} + v_j + w_k.$$

The same weakly informative priors selected for the separate models, described in the design analysis (section 5.2.1), were used for the population-level effects, group-level effects, and covariance parameters. The model was estimated using six chains with 8000 iterations each, half of which were discarded as burn-in. To pre-empt divergent transitions, ‘adapt_delta’ was set to 0.99. The model appeared to converge well under this specification. The full results of the joint model are presented in Table 5.1.

Table 5.1

Bayesian Multilevel Logistic Model of Response Accuracy, including both Global and Local Judgements, in Study 3

Parameter	Estimate[a]	Error[b]	95% CI[c]	Eff Sample
Population-level Effects				
Intercept	0.68	0.25	[0.19, 1.17]	6851
Global perceived comprehension	0.05	0.04	[-0.02, 0.13]	19022
Local perceived comprehension	0.17	0.03	[0.11, 0.23]	25281
Group-Level Variance				
Participant (intercept)	0.72	0.05	[0.64, 0.82]	6155
Participant (global perceived comprehension)	0.07	0.04	[0.00, 0.16]	4206
Participant (local perceived comprehension)	0.06	0.04	[0.00, 0.14]	6911
Text (intercept)	0.69	0.26	[0.24, 1.26]	2885
Question (intercept)	1.17	0.11	[0.97, 1.40]	4167
Covariance of intercept and slope variance				
Intercept, global perceived comprehension	0.30	0.39	[-0.63, 0.90]	18075
Intercept, local perceived comprehension	-0.32	0.40	[-0.92, 0.62]	18900
Global perceived comprehension, local perceived comprehension	-0.14	0.49	[-0.91, 0.82]	8471

Note: Population-level effect estimates are presented in logits. Rhat values for all parameters = 1.00. CI = credible interval. Eff Sample = number of effective samples, obtained using the bayestestR package (Makowski et al., 2019).

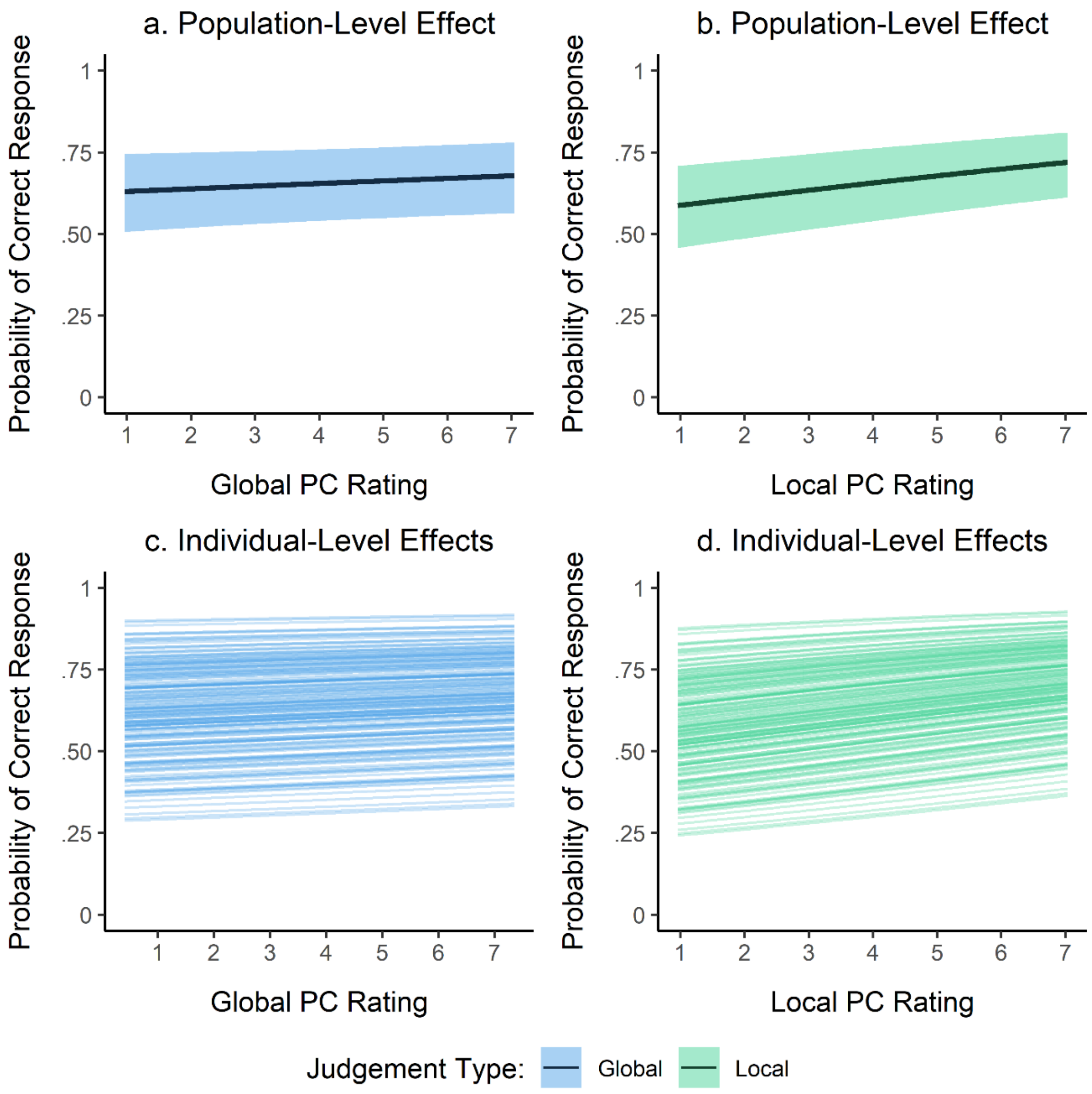
[a]Estimate refers to the mean of the marginal posterior distribution of the parameter. [b]Error refers to the standard deviation of the marginal posterior distribution of the parameter. [c]Credible intervals represent the upper and lower values within which 95% of the estimated parameter values in the posterior distribution are contained.

The means of the posterior distributions for both global and local judgements were estimated to be positive, indicating that higher ratings of perceived comprehension at the text-level and at the idea-level increase the log odds of answering a comprehension question correctly. For each unit increase in perceived comprehension, the log odds of a correct response was estimated to increase by 0.05, given a global judgement, and 0.17, given a local judgement. However, the effect of global judgements should be interpreted with caution, as the 95% credible interval overlaps with zero, with negative values associated with some posterior plausibility (see Table 5.1).

Based on model-fitted predictions of the marginal effects, the lowest and highest global perceived comprehension judgements would correspond to an expected probability of a correct response of 0.63 (*SE* = 0.06) and 0.68 (*SE* = 0.05), respectively. In contrast, for local perceived comprehension judgements, these expected probabilities are 0.58 (*SE* = 0.06) and 0.71 (*SE* = 0.05), respectively. The uncertainty in both of the effects can be seen in Figure 5.6, which shows the model-fitted predictions of correctly answering a comprehension question given a global (5.6a) or local (5.6b) judgement.

Figure 5.6

Estimated Effect of Perceived Comprehension, at the Population- and Individual-Level, by Judgement Type, Given a Joint Judgement Model, in Study 3



Note. Marginal model-fitted predictions at the population-level (a. and b.) and conditional model-fitted predictions at the individual-level (c. and d.) are plotted, coloured according to whether the judgement of perceived comprehension was made at the level of the text (blue) or the level of information within a text (green). The judgement type not plotted in each plot, respectively, is held at its mean. Panels a. and b. show the expected probability of a correct response for the 'average' participant, responding to the 'average' question concerning an 'average' text, with the 95% credible interval shaded. Panels c. and d. show the expected probability of a correct response for each participant, given an ‘average’ question concerning an ‘average’ text. PC = perceived comprehension.

Limited participant-level variability was estimated in the effect of perceived comprehension for both global and local judgements. Participant-level model-fitted estimates of the effect of perceived comprehension are shown in Figure 5.6c, for global judgements, and Figure 5.6d, for local judgements. As can be seen in these plots, the predictive relationship between perceived and assessed comprehension remained weakly positive across participants, for both global and local judgements.

To compare the relative strength of the predictive relationship for global and local perceived comprehension judgements in the joint model, the 80% HDI for each effect was calculated. Non-overlapping 80% HDIs were estimated: global judgment 80% HDI = [0.01, 0.10]; local judgement 80% HDI = [0.13, 0.21]. It should be noted, however, that the 95% credible intervals for these parameters do overlap (see Table 5.1). Given the symmetrical posterior distributions for these effects, the 95% HDIs also overlap. Further examination of the posterior distributions indicated that HDIs of up to 90% provide non-overlapping intervals for these effects.

Separate Models. The separate judgement models, defined in (11) and (12), repeated below for convenience, were fitted using a Bayesian estimation framework:

$$\log\left(\frac{P_{ijk}}{1-P_{ijk}}\right) = \beta_0 + (\beta_1 + u_{1i})GlobalJudgement_{ij} + u_{0i} + v_j + w_k, \qquad (11)$$

$$\log\left(\frac{P_{ijk}}{1-P_{ijk}}\right) = \beta_0 + (\beta_2 + u_{2i})LocalJudgement_{ijk} + u_{0i} + v_j + w_k. \qquad (12)$$

Observations of assessed comprehension, from individual i on text j responding to question k, were used at the outcome measure Y_{ijk}. Judgements of perceived comprehension, elicited in response to either global or local judgement prompts, standardised prior to model fitting, were separately entered as predictors of the log odds of correctly responding to the comprehension question. Group-level variance in the intercept of both models was estimated for participants u_{0i}, texts v_j and questions w_k. Participant-level variability in the effect of global u_{1i} and local u_{2i} perceived comprehension judgements was also included in each model. The same weakly informative priors described in the design analysis (section 5.2.1) were used for population-level effects, group-level effects, and covariance parameters in the separate models.

The model was estimated using six chains with 8000 iterations each, half of which were discarded as burn-in. To pre-empt divergent transitions, 'adapt_delta' was set to 0.99. The model appeared to converge well under this specification. The full results of the separate models are presented in Table 5.2 (global judgement model) and Table 5.3 (local judgement model).

Table 5.2

Bayesian Multilevel Logistic Model of Response Accuracy, including only Global Judgements, in Study 3

Parameter	Estimate[a]	Error[b]	95% CI[c]	Eff Sample
Population-level Effects				
Intercept	0.68	0.25	[0.17, 1.18]	8847
Global perceived comprehension	0.14	0.03	[0.08, 0.21]	26807
Group-Level Variance				
Participant (intercept)	0.72	0.05	[0.64, 0.82]	7730
Participant (global perceived comprehension)	0.06	0.04	[0.00, 0.16]	4362
Text (intercept)	0.69	0.26	[0.23, 1.27]	3427
Question (intercept)	1.17	0.11	[0.97, 1.41]	5948
Covariance of intercept and slope variance				
Intercept, global perceived comprehension	0.20	0.45	[-0.80, 0.94]	22532

Note: Population-level effect estimates are presented in logits. Rhat values for all parameters = 1.00. CI = credible interval. Eff Sample = number of effective samples, obtained using the bayestestR package (Makowski et al., 2019).

[a]Estimate refers to the mean of the marginal posterior distribution of the parameter. [b]Error refers to the standard deviation of the marginal posterior distribution of the parameter. [c]Credible intervals represent the upper and lower values within which 95% of the estimated parameter values in the posterior distribution are contained.

Table 5.3

Bayesian Multilevel Logistic Model of Response Accuracy, including only Local Judgements, in Study 3

Parameter	Estimate[a]	Error[b]	95% CI[c]	Eff Sample
Population-level Effects				
Intercept	0.67	0.26	[0.17, 1.18]	5915
Local perceived comprehension	0.19	0.03	[0.13, 0.24]	23201
Group-Level Variance				
Participant (intercept)	0.73	0.05	[0.64, 0.82]	5973
Participant (local perceived comprehension)	0.05	0.04	[0.00, 0.13]	5657
Text (intercept)	0.70	0.26	[0.25, 1.29]	3084
Question (intercept)	1.16	0.11	[0.97, 1.40]	4598
Covariance of intercept and slope variance				
Intercept, local perceived comprehension	-0.32	0.44	[-0.96, 0.74]	19180

Note: Population-level effect estimates are presented in logits. Rhat values for all parameters = 1.00. CI = credible interval. Eff Sample = number of effective samples, obtained using the bayestestR package (Makowski et al., 2019).

[a]Estimate refers to the mean of the marginal posterior distribution of the parameter. [b]Error refers to the standard deviation of the marginal posterior distribution of the parameter. [c]Credible intervals represent

the upper and lower values within which 95% of the estimated parameter values in the posterior distribution are contained.

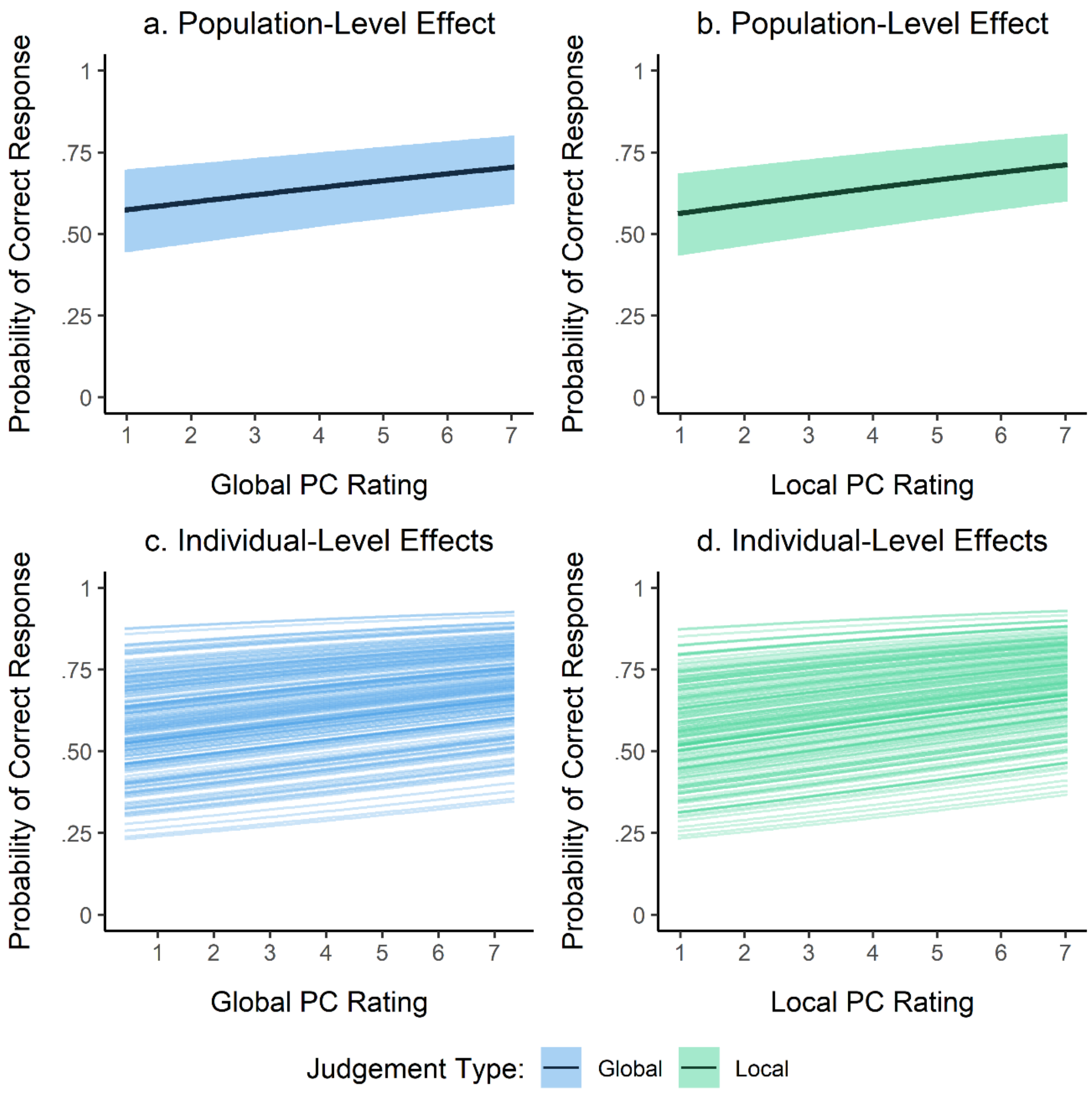
In each of the separate models, the posterior means for global and local judgements were estimated to be positive, indicating that higher ratings of perceived comprehension at both the text-level and the idea-level increase the log odds of answering a comprehension question correctly. Estimated posterior means were larger than those obtained from the model defined in (13), particularly for the global judgement effect. For each unit increase in perceived comprehension, the log odds of a correct response was estimated to increase by 0.14, given a global judgement, and 0.19, given a local judgement. For both global and local judgements, the 95% confidence interval for the effect contained only positive values (see Table 5.2 and 5.3).

Based on model-fitted predictions of the marginal effect, the lowest and highest global perceived comprehension judgements corresponded to an estimated probability of a correct response of 0.57 (*SE* = 0.06) and 0.70 (*SE* = 0.05), respectively, illustrated in Figure 5.7a. For local perceived comprehension judgements, these probabilities were estimated to be 0.56 (*SE* = 0.06) and 0.71 (*SE* = 0.05), given the lowest and highest local perceived comprehension judgements respectively, shown in Figure 5.7b.

Limited participant-level variability was observed in the effect of perceived comprehension for both global and local judgements. The participant-level model-fitted predictions are shown in Figure 5.7c, for global judgements, and Figure 5.7d, for local judgements. As illustrated in these figures, the predictive relationship between perceived and assessed comprehension remains comparable in magnitude, for both global and local judgements, across participants.

Figure 5.7

Estimated Effect of Perceived Comprehension, at the Population- and Individual-Level, by Judgement Type, Given Separate Judgement Models, in Study 3



Note. Marginal model-fitted predictions at the population-level (a. and b.) and conditional model-fitted predictions at the individual-level (c. and d.) are plotted, coloured according to whether the judgement of perceived comprehension was made at the level of the text (blue) or the level of information within a text (green). Panels a. and b. show the expected probability of a correct response for the ‘average’ participant, responding to the ‘average’ question concerning an ‘average’ text, with the 95% credible interval shaded. Panels c. and d. show the expected probability of a correct response for each participant, given an ‘average’ question concerning an ‘average’ text. PC = perceived comprehension.

To compare the relative predictive accuracy of model (11) and (12), the difference in the ELPD was calculated (Vehtari et al., 2016; Watanabe, 2010). A difference of -14.1 in the ELPD was estimated, in favour of the local judgement model: global judgement model (11) $\widehat{ELPD}$ = -6972.74; local judgement model (12) $\widehat{ELPD}$ = -6958.61). The standard error of the difference was 6.37. Therefore, the interval defined by the difference plus or minus three times this error was [-33.24, 5.00]. While the ELPD difference was estimated to be greater than four, the interval for the difference spanned zero. Therefore, the criteria defined in the preregistration to evidence greater predictive utility was only partially met. However, the 95% confidence interval, a more frequently selected probability for intervals, for this difference may also be reasonably considered, assuming the standard error of the ELPD difference is normally distributed (determined by factors including model misspecification, the number of observations and size of ELPD difference, Sivula et al., 2020). The 95% confidence interval for this estimated ELPD difference would exclude zero: 1.96 multiplied by the standard error of the ELPD yields the interval [-26.62, -1.63], indicating greater predictive accuracy in the local judgement model (12).

5.3.3 Sensitivity Analysis

A sensitivity analysis was conducted to explore the robustness of the findings of the main analysis (section 5.3.2) to alternative analytical choices. In this section, the alternative choices examined are first discussed. An outline of the motivation is then followed by an evaluation of the impact on the estimated effects of interest and a comparison of the predictive accuracy of the model specifications. Lastly, across the specifications, support is evaluated for i) a difference between the effect magnitudes estimated in the joint models and ii) a difference in the predictive accuracy of the separate models.

Alternative Model Specifications. Five analytic choices were identified as potentially impacting the findings of Study 3: i) the experimental design used to elicit both global and local judgements from participants, ii) the assumed shape of the predictive association between perceived and assessed comprehension, iii) the impact of multiple-choice questions options on the question response process, iv) the chosen priors for the effects of interest, and v) the use of the participant exclusion criteria. Each of these are discussed in depth below. To limit repetition, where the motivation and implementation of an alternative specification is equivalent to that outlined previously, a brief description is given with reference to the fuller discussion.

Experimental Design. In Study 3, a within-participants experimental design was used, with the same participants providing both global and local perceived comprehension judgements. Due to this repeated-measures design, participants' internal global judgement processes may have differed from Study 1 and 2. For example, metacognitive experiences generated from reflecting on the understanding of the information targeted by the local judgement prompts may have been used to inform global judgements in Study 3 (Dunlosky et al., 2005; Griffin, Wiley & Thiede, 2019). Such interference in the global judgement process, due to simultaneously providing local judgements, would complicate the comparison of the predictive relationship between these judgment types and assessed comprehension.

To avoid this potential issue, a between-participants design could be used to elicit global judgements independent of local judgements. While Study 1 and 2 indicated limited individual variability in the strength of the association between global judgements of perceived comprehension and assessed comprehension, the process by which individuals make metacomprehension judgements and the factors which influence predictive accuracy at an individual level remain poorly understood. As a result, it is unclear whether the average, population-level predictive accuracy, for both global and local judgements, is likely to be well captured at the sample level. Any observed difference in predictive magnitude between judgement types, therefore, may be attributed to individual differences alone.

To partially address these issues, the responses collected in Study 2, consisting of global judgements only, can be analysed alongside the local judgements obtained in Study 3. This was considered appropriate given that the studies were fundamentally similar. The same stimulus texts were used in the metacomprehension tasks and, while 11 questions were altered between studies, the comprehension questions assessed understanding of the same information within texts. Further, although the judgement prompt used to elicit global judgements differed between studies ("How much of the text do you understand?" and "Overall, how well do you understand the text?"), the sensitivity analyses in Study 1 and 2 suggested these prompts elicit similar responses. Combining the data obtained in both Study 2 and 3, therefore, offered a cogent method to evaluate the influence of simultaneously eliciting both global and local judgements of comprehension and the resulting impact on the estimated predictive relationship between perceived and assessed comprehension between judgement types.

The model defined in (13) was altered to include global judgements from Study 2 and local judgements from Study 3: a binary indicator variable was used to capture whether judgements corresponded to Study 2 ($Study_{ijk} = 0$) or Study 3 ($Study_{ijk} = 1$). The study indicator was entered as a predictor alongside perceived comprehension judgements and the interaction between these two variables. Three population-level effects were estimated: the effect of the perceived comprehension judgement in Study 2 (β_1), the effect of the study (β_2), and the change (interaction) in the effect of perceived comprehension for Study 3 (β_3). Participant-level variability in the effect of the perceived comprehension (u_{1i}) was estimated. Furthermore, participants (I = 396) and questions (K = 89) were recoded with unique identifiers. The model fitted was, therefore:

$$\log\left(\frac{P_{ijk}}{1-P_{ijk}}\right) = \beta_0 + (\beta_1 + u_{1i})Judgement_{ij} + (\beta_2 + v_{1j} + w_{1k})Study_{ijk}$$

$$+ \beta_3 Judgement_{ij}Study_{ijk} + u_{0i} + v_{0j} + w_{0k}\,.$$

For the separate judgement models, the model defined in (11) was fit to the responses collected in Study 2. For this model, observations of $GlobalJudgement_{ij}$ were those obtained in Study 2. The model defined in (12) was not altered and was fit to local judgements obtained in Study 3.

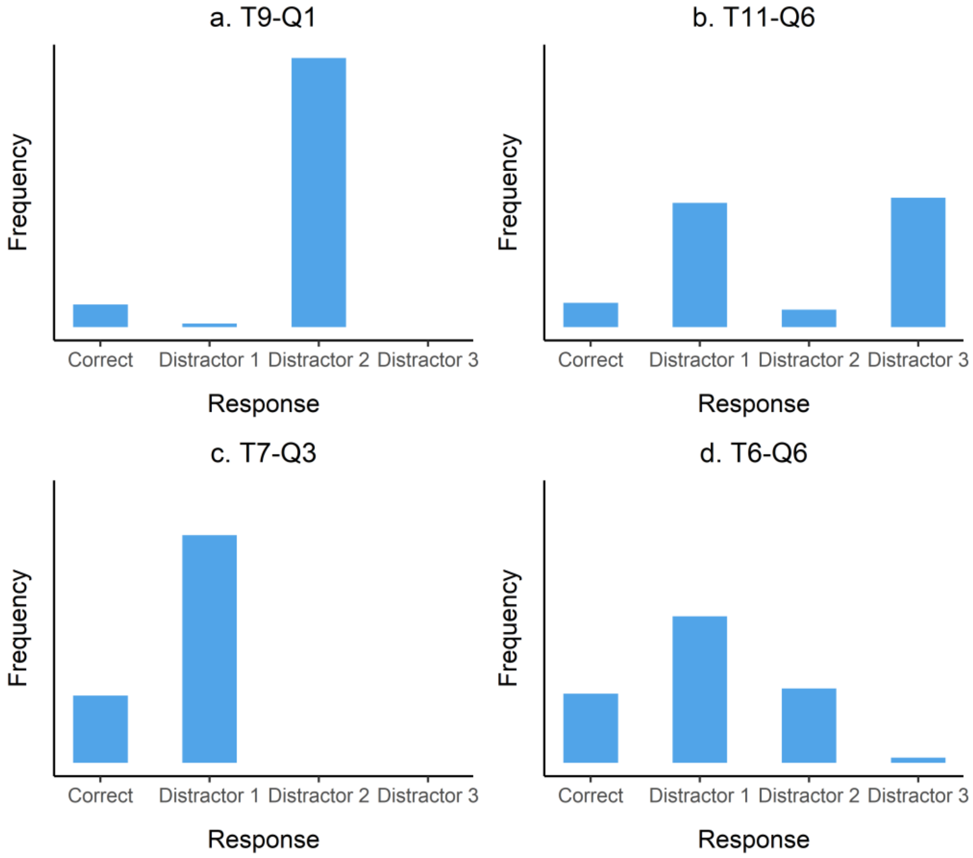
Nonlinear Relationship. The models defined in (11), (12), and (13) assumed that the relationship between both global and local judgements of perceived comprehension and assessed comprehension was linear on the logit scale and that judgements could be treated as interval-level data. As discussed previously (Chapter 3; section 3.3.3), such an assumption, if incorrect, can lead to inaccurate estimates of the predictive relationship. Instead, metacomprehension judgements and their relationship with assessed comprehension may be better captured by monotonic model (Bürkner and Charpentier, 2020). Models (11), (12), and (13) were altered, therefore, to estimate the monotonic effect of perceived comprehension. Six simplex parameters (see Chapter 3; section 3.3.3) were estimated, capturing the expected difference in the probability of a correct response between adjacent ratings of comprehension. For both the joint and separate judgement models, estimates were obtained for the expected difference between adjacent ratings of perceived comprehension and the overall average expected difference in probability of observing a correct response between ratings.

Random Guessing. In the metacomprehension task, each of the six comprehension questions per text were presented with multiple response options. As discussed previously (Chapter 3; section 3.3.3), closed-form response options can introduce bias in the estimation of the relationship between perceived and assessed comprehension (Vuorre & Metcalfe, 2022). A model in which the intercept is constrained can be used to incorporate a random guessing response process, estimating the influence of the covariates in addition to some fixed probability of a correct response due to chance alone (Bürkner, 2022). However, the assumption of a random response process may not be accurate (Higham, 2007; Vuorre & Metcalfe, 2022).

The selection of response options in Study 3 was considered to evaluate whether participants response behaviour appears consistent with a random guessing response process. Figure 5.8, showing responses to four questions with the lowest response accuracy rates, suggests that observed response selection is inconsistent with a random selection process.

Figure 5.8

Responses to Four Comprehension Questions with the Lowest Response Accuracy in Study 3



Note. T = text. Q = question.

Nevertheless, participants may engage in guessing processes to some extent along side knowledge-drive selection procedures (e.g., Embretson & Weztel, 1987). To evaluate the bias that is introduced in the estimated relationship between perceived and assessed comprehension if random response selection occurs but is not accounted for, therefore, the models defined in (11), (12), and (13) were altered to incorporate a constrained intercept.

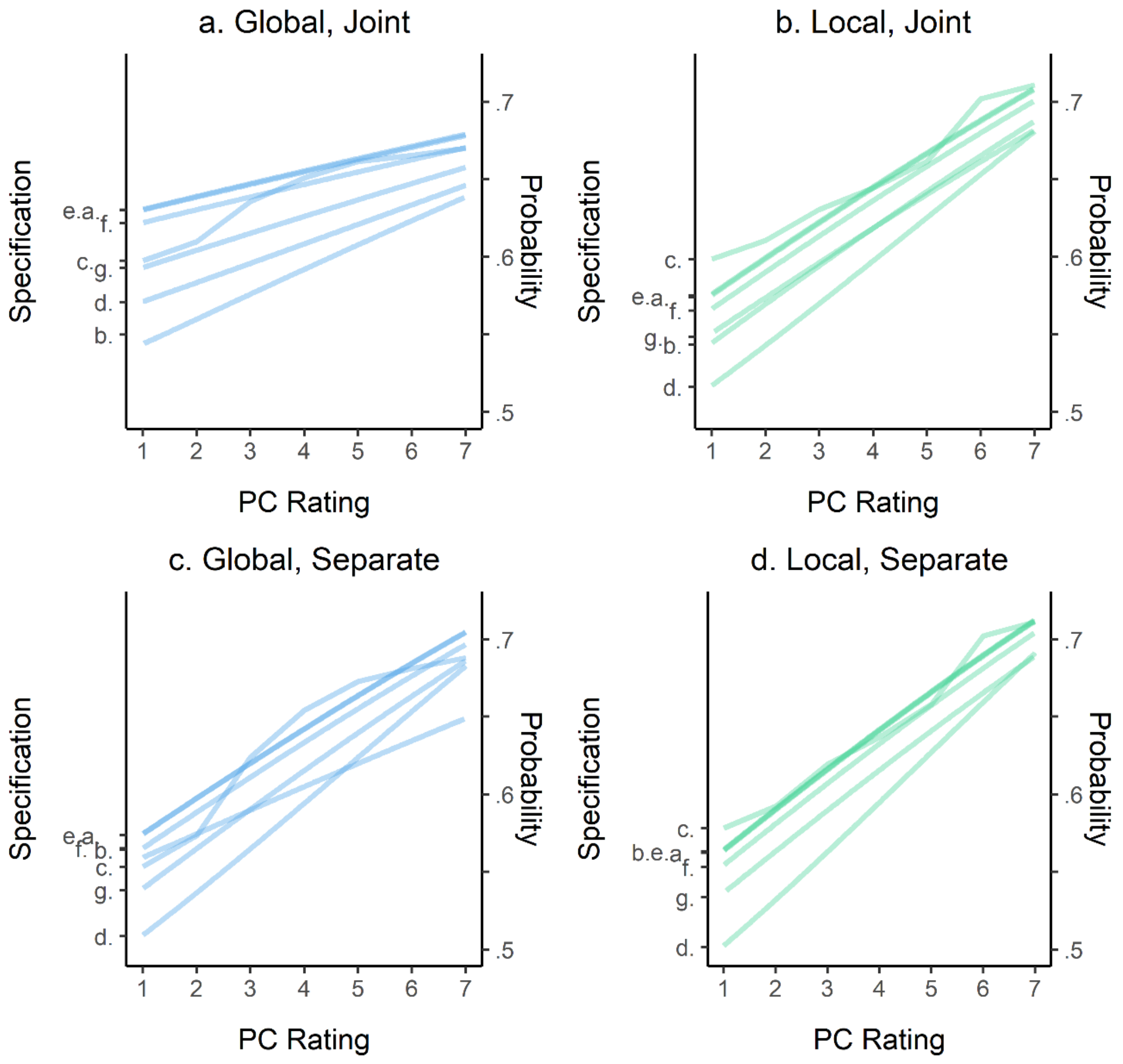
Variation in Priors. Weakly informative priors were chosen for model parameters in the main analysis (section 5.3.2). The extent to which the posterior distributions for the effect parameters are robust to reasonable variation in the prior specification for effect parameters was examined to evaluate the impact of the choice of priors. The models defined in (11), (12), and (13) were refit with looser and tighter priors, as described in the sensitivity analysis of Study 1 (Chapter 3; section 3.3.3).

No Exclusions. Consistent with Study 2, participants were excluded from the main analysis if any reading times for the 13 health-related texts were faster than a reading rate of 300 wpm. The motivation for excluding participants and the potential issues with a binary decision threshold for exclusion are discussed in the sensitivity analysis of Study 1 (Chapter 3; section 3.3.3). To evaluate whether the estimated effects were influenced by participant exclusion, the models defined in (11), (12), and (13) were refitted using the full dataset without exclusions (*I* = 197).

Model Comparison: Effects of Interest. In models jointly including global and local judgements, as in (13), and models separately including global or local judgements, as in (11) and (12), respectively, alternative specifications predominantly produced similar effect estimates. For both judgement types, the posterior means for the effect of judgements on the probability of answering a comprehension question correctly remained positive in all specifications. Further, limited participant-level variability in the predictive relationship between judgements and assessed comprehension was consistently found. The model-fitted, predicted probability of correctly answering a comprehension question for each rating response, across all specifications, are shown in Figure 5.9. For models featuring both global and local judgements, the effect of each unit increase in global judgements, holding local judgements at the mean, is shown in Figure 5.9a, and the effect of each unit increase in local judgements, holding global judgements at the mean is shown in Figure 5.9b. Figures 5.9c and 5.9d show the effects of each unit increase in global judgements and local judgements, respectively, when these judgements are specified within separate models.

Figure 5.9

Estimated Effect of Perceived Comprehension for Each Model Specification in Study 3



Note. Marginal model-fitted predictions at the population-level for model specifications including both judgements (joint model) made at the level of the text (a.) and information within a text (b.), and for model specifications including either judgements (separate models) made at the level of a text (c.)

or information within a text (d.), as predictors are plotted. Model specifications a-g refer to: a. main analysis model, b. between-participants design, c. monotonic effect, d. constrained intercept, e. looser effect priors, f. tighter effect priors, g. no exclusion criteria. PC = perceived comprehension.

With respect to global judgements of comprehension, two specifications notably influenced the estimated relationship between perceived and assessed comprehension. Firstly, modelling the relationship as monotonic (labelled as specification c.), resulted in clear non-linearity, with a larger increase in the probability of a correct response was estimated between the second and third ratings on the judgement scale. Nonlinearity in the relationship was greater when global judgements were modelled separately to local judgements. Secondly, using the global judgements collected in Study 2 produced an estimated magnitude for the effect which was similar to that estimated in the joint model using Study 3 data. However, given a separate judgement model, the magnitude was comparably lower to that estimated using Study 3 data ($\hat{\beta}_1$= 0.10, 95% CI = [0.04, 0.16]).

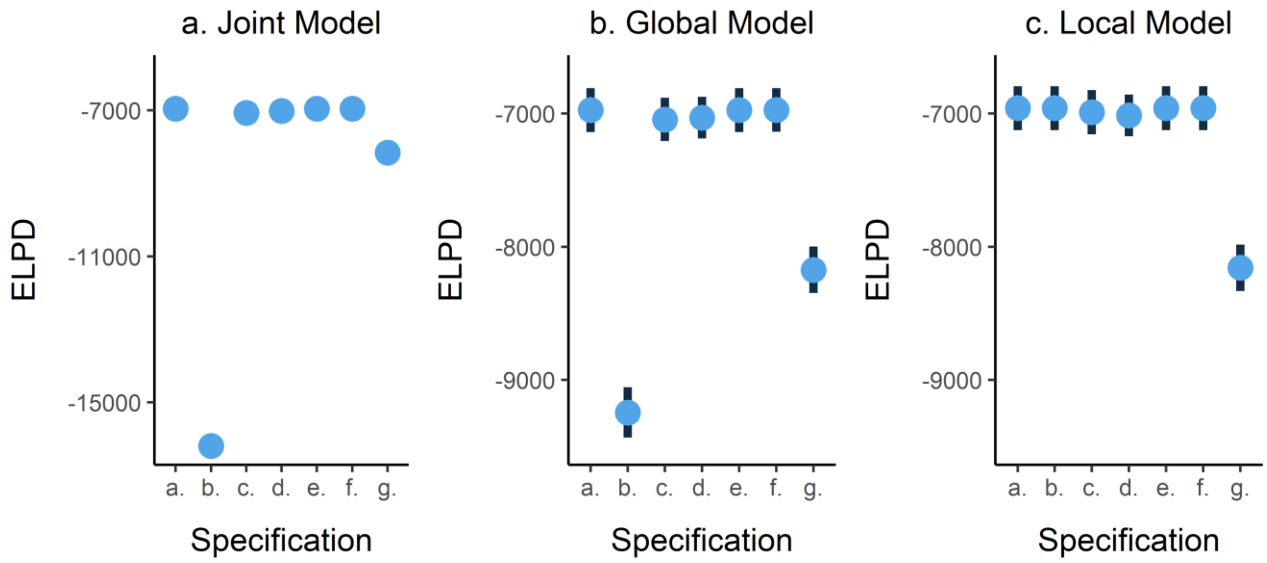
With respect to local judgements, very limited differences were observed across alternative model specifications. Modelling the relationship as monotonic (labelled as specification c.), produced a larger increase in the expected probability of a correct response between the fifth and sixth highest ratings on the judgement scale. In addition, modelling a constrained intercept to account for a random guessing response process (labelled as specification d.) resulted in a stronger estimated relationship between perceived and assessed comprehension, particularly when local judgements were modelled separately to global judgements.

Model Comparison: Predictive Accuracy. The relative predictive accuracy of the model specifications was compared to evaluate whether predictive performance differed across the analytic choices considered. As in Chapter 3 (section 3.3.3), predictive performance was evaluated in two ways for each model: i) calculating the estimated expected log predictive density (ELPD) and ii) posterior predictive checks. The ELPD was calculated using the Widely Applicable Information Criterion (WAIC) computation (Vehtari et al., 2016; Watanabe, 2010), to provide a measure of out-of-sample predictive accuracy. Posterior predictive accuracy was evaluated by calculating the discrepancy between observed response accuracy and model predictions, to provide a measure of within-sample predictive accuracy (Gabry et al., 2019).

The estimated ELPDs for each specification are shown in Figure 5.10, for the joint judgement models (5.10a), global judgement models (5.10b), and local judgement models (5.10c). Note that direct comparisons between ELPDs estimated from models fitted to different sized datasets (here, specification b. and g., referring to models fitted using Study 2 data and no exclusion criteria, respectively) are not appropriate, given that possible differences in predictive accuracy are indistinguishable from the impact of additional observations.

Figure 5.10

ELPD Estimates for Each Model Specification in Study 3



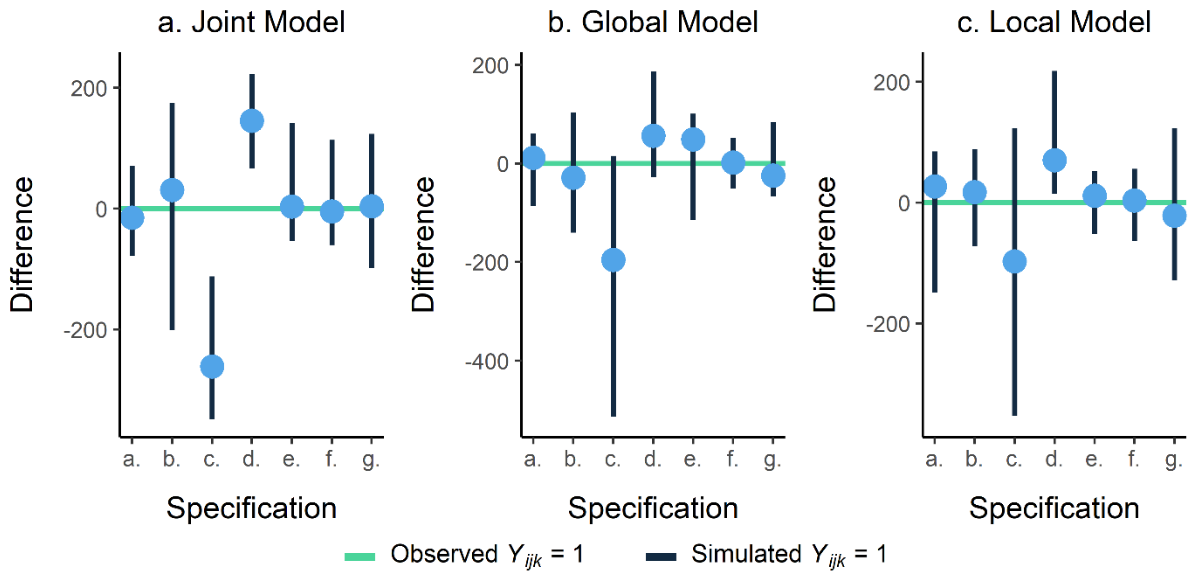
Note. Points (blue circles) show the estimated ELPD and bars show three times the standard error of the estimate, for model specifications including (a.) both judgements made at the level of the text and information within a text (joint model), (b.) judgements made only at the level of a text (global model), and (c.) judgements made only at the level of information within a text (local model) as predictors. No error bars can be seen in the joint model plot due to the range of the horizontal axis, however, error bars are approximately equivalent to those shown in the global model plot. Model specifications a-g refer to: a. main analysis model, b. between-participants design, c. monotonic effect, d. constrained intercept, e. looser effect priors, f. tighter effect priors, g. no exclusion criteria. ELPD = expected log predictive density.

Across the specifications fitted to the same sized datasets (a., and c. to f.), differences in the estimated ELPD were minimal and were located within the intervals defined by three times the standard errors of the ELPD estimates, indicating highly similar out-of-sample predictive accuracy.

Within-sample predictive accuracy was broadly similar across the alternative specifications. Figure 5.11 shows the estimated posterior predictive accuracy for each specification, for the joint judgement models (5.11a), global judgement models (5.11b), and local judgement models (5.11c). In the majority of specifications, the difference between observed and simulated response accuracy appeared limited and unbiased. However, posterior predictions from models which assumed a monotonic relationship between perceived and assessed comprehension (labelled as specification c.) tended to considerably underestimate correct responses. In addition, posterior predictions from models with a constrained intercept (labelled as specification d.) tended to overestimate correct responses, particularly in the joint judgement model. Further, while generally unbiased, given a joint judgement model, using Study 3 and Study 2 data (labelled as specification b.) produced notable variation in posterior predictive accuracy.

Figure 5.11

Posterior Predictive Accuracy for Each Model Specification in Study 3



Note. Plotted according to model specifications including (a.) both judgements made at the level of the text and information within a text (joint model), (b.) judgements made only at the level of a text (global model), and (c.) judgements made only at the level of information within a text (local model) as predictors. Light blue circles show the mean difference between observed correct responses ($Y_{ijk} =$ 1) and simulated correct responses. Dark blue lines show the range of differences estimated between observed and simulated correct responses. Model specifications a-g refer to: a. main analysis model, b. between-participants design, c. monotonic effect, d. constrained intercept, e. looser effect priors, f. tighter effect priors, g. no exclusion criteria.

Model Comparison: Evidence of a Difference. To aid in the comparison of whether judgements of comprehension made at the level of a text or at the level of information within a text are better predictors of the understanding, the relative predictive performance of these judgement types was evaluated across the alternative model specifications. Evidence in support of a difference in predictive performance was defined in two ways: i) non-overlapping 80% HDIs for the effects in the joint model and ii) a difference of greater than four in the estimated ELPDs of the separate judgement modes, with the interval for this difference (defined by ±3 standard errors) excluding zero. For each specification considered, these values are illustrated in Figure 5.12. The left column of plots in Figure 5.12 shows the posterior distributions of the population-level judgement effects, with the 80% HDI shaded blue and green, for global and local judgements, respectively. The right column of plots in Figure 5.12 illustrates the estimated ELPD difference relative to zero, shown as a dashed line, with error bars capturing three times the standard error. Note that no ELPD difference plot is shown for specification b., due to differences in the size of the dataset used in the global judgement model (using Study 2 observations).

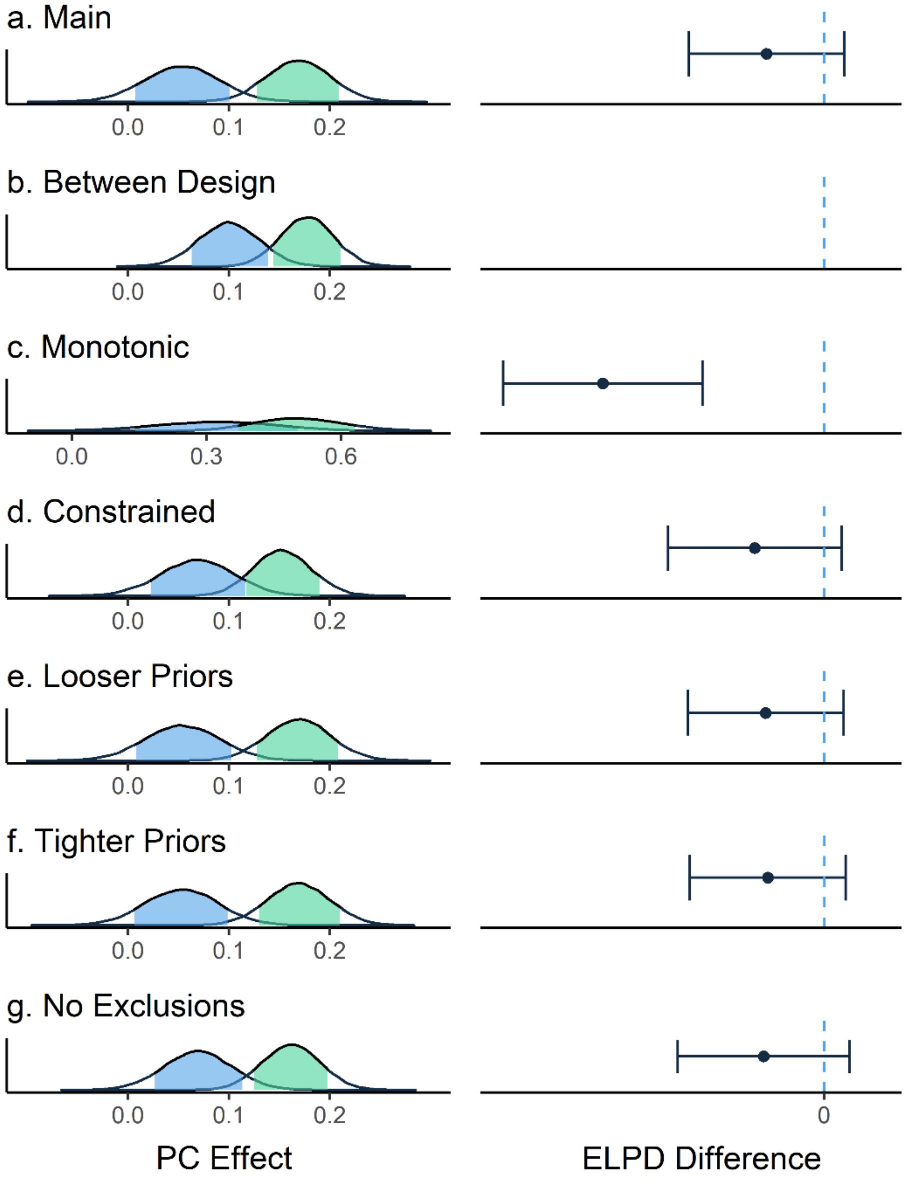
Each of the alternative model specifications produced 80% HDIs for the effects of global and local judgements located more closely together than in the main analysis. This was particularly true for the model assuming a random guessing response process, labelled d. Considerable overlap in the 80% HDIs for the specification assuming a monotonic relationship between perceived and assessed comprehension, labelled as specification c., can be seen in Figure 5.12. This overlap largely reflects the high uncertainty in the posterior distributions of the parameters plotted (the average expected difference between adjacent ratings). Given the difference in the nature of the parameter plotted for this model, this contrast is somewhat less informative. Nevertheless, all specifications estimating a linear relationship between perceived and assessed comprehension yield 80% HDIs for global and local judgements which do not overlap.

Consistent with the main analysis, the interval defined by the estimated ELPD difference plus or minus three times the standard error of the difference overlaps with zero in the majority of alternative specifications. However, as can be seen in Figure 5.12, assuming a monotonic relationship between global and local judgements of perceived comprehension and understanding produced an interval for the ELPD difference between models, in favour of the local judgement model, which did not overlap with zero (difference in $\widehat{ELPD}$ = -54.26, [-

78.72, -29.81]). Overall, while alternative specifications similarly yield an estimated ELPD difference of greater than four in the direction of greater predictive performance of local judgements of perceived comprehension, the interval for this difference did not meet the predefined criteria.

Figure 5.12

Evidence for a Difference Between Global and Local Judgements Across Model Specifications in Study 3



Note. The left column shows the posterior distributions of the estimated effects of perceived comprehension, from model specifications including both judgements made at the level of a text (global judgement) and information within a text (local judgement) as predictors. The 80% highest density interval of the posterior distribution for global and local judgements are shaded in blue and green, respectively, For the monotonic model (c.), the posterior distributions of the average expected difference between adjacent ratings are plotted. The right column shows the estimated difference in the ELPD between model specifications including either judgements made at the level of a text or information within a text as predictors. Negative differences (to the left of zero) correspond to a lower ELPD for the local judgement model, while positive differences (to the right of zero) would correspond to a lower ELPD for the global judgement model. No ELPD difference plot is shown for specification b. due to differences in the size of the dataset these models were fitted to. PC = perceived comprehension, ELPD = expected log predictive density.

5.4 Discussion

Study 3 was conducted to address RQ5 (see section 5.1.2). In this section, the main findings are first discussed with respect to previous research, followed by a consideration of the theoretical implications of the results and the limitations of this research.

5.4.1 Main Findings

In the model which included both global and local judgements as predictors (joint model), evidence was observed which supported a difference in the predictive relationship between local and global judgements and assessed comprehension. Within this model, non-overlapping 80% HDIs for the effects of global and local judgements were found. When the relative contributions of global and local judgements are evaluated jointly within a single model, therefore, local judgements of perceived comprehension had a greater influence on the probability of observing a correct response than global judgements. This finding was replicated in all models considered in the sensitivity analysis with the exception of the monotonic model, owing to the considerably higher uncertainty in the posterior distributions of the parameters capturing the estimated average difference between adjacent ratings. This indicates that local metacomprehension judgements provide comparably greater insight into comprehension outcomes, with global judgements contributing little additional information beyond this.

Partial support for a difference in the predictive relationship between judgement types was observed in the models which separately specified global and local judgements as predictors (separate models). In comparing the relative predictive accuracy of these models, an estimated difference in the ELPD of magnitude greater than four, in favour of the local judgement model, was found. However, the interval constructed from this value plus or minus three times the standard error of the estimate overlapped with zero. This suggests that using either global or local judgements of perceived comprehension as predictors of correct responses may result in limited differences in the capacity to accurately predict comprehension outcomes. However, this conclusion is influenced by the conservative choice for the ELPD difference interval. Considering, instead, the 95% confidence interval for the ELPD difference (1.96 times the standard error), this interval did not overlap with zero, indicating that on 95% of occasions, the estimated ELPD difference will be a non-zero value, in favour of the local judgement model. This result was replicated in all models considered in the sensitivity analysis with the exception of the monotonic model, for which the interval constructed from three times the standard error also did not overlap with zero.

Considering the evidence in support of a difference between global and local judgements observed in the joint model, and the partial support observed in the separate models, on balance, the results suggest that local judgements are likely a better predictor of assessed comprehension than global judgements. Eliciting judgements of perceived comprehension at the level of specific pieces of information, which are subsequently evaluated for understanding, likely provides a more informative measure of future comprehension performance than judgements elicited at the level of a text. Specifically, unit changes in local judgements were estimated to have between two to three times the effect of global judgements on the log odds of observing a correct response on the comprehension task, based on posterior means of the main analysis models. This finding is consistent with research which has shown that judgements made in respect of a smaller quantity of information associate more strongly with performance than judgements made across a whole text (Dunlosky et al., 2005). In contrast with previous research, however, no pre-judgement retrieval attempt was required to produce the observed difference and no test expectations were made salient at the time of eliciting metacomprehension judgements (Dunlosky, Rawson & McDonald, 2002; Dunlosky et al., 2005). The results of Study 3, therefore, generalise more readily to the context within which reader panellists provide judgements of text comprehensibility.

While local judgements were found likely to predict comprehension outcomes to a greater extent than global judgements, neither type of judgement was found to have a strong predictive relationship with assessed comprehension. In fact, estimates obtained for the effect of local judgements remained lower than those observed for the effects of reading skill and background knowledge in Study 1. This suggests that individual difference measures may be more predictive of comprehension outcomes than metacomprehension judgements. This finding is consistent with research which shows metacomprehension judgements are only weakly positively associated with comprehension, on average (Dunlosky & Lipko, 2007; Lin & Zabrucky, 1998; Maki, 1998; Prinz et al., 2020a; Yang et al., 2022). However, in contrast to research suggesting that metacomprehension accuracy varies considerably across individuals (Chiang et al., 2010; Glenberg & Epstein, 1985; Jee et al., 2006), limited individual variability in the magnitude of the predictive relationship between both global and local judgements and assessed comprehension was observed. Overall, therefore, while readers show limited predictive accuracy in discriminating between texts and specific ideas within texts which are more or less well understood, both types of judgement positively associate with comprehension outcomes across individuals.

In addition to these main findings, the results of the sensitivity analysis suggest that the predictiveness of global judgements may be influenced by concurrently eliciting local judgements. In an attempt to evaluate the potential impact of the repeated measures design on global and local judgements, global judgements of comprehension observed in Study 2 were used to provide an approximate design contrast. As a result of the shared variance between global and local judgements observed in Study 3 (see Figure 5.5), within the joint model, global judgements made in Study 3 contributed less unique variance to the outcome than local judgements. Removing local judgements from this model resulted in additional variance in the outcome attributed to global judgements. This relationship is not observed when global judgements are elicited in the absence of local judgements, as indicated by the models fit to Study 2 data. This suggests that eliciting local judgements prior to global judgements of comprehension could increase the shared variance between these judgement types, slightly improving the predictive accuracy of global judgements as a result. However, as this was not the main focus of the present study, further research is required to robustly evaluate this relationship. Furthermore, regardless of such a potential increase in predictiveness of global judgements, the results suggest that local judgements remain better predictors of assessed comprehension.

5.4.2 Theoretical Implications

Overall, the findings of this study indicate that local judgements of comprehension are likely more predictive of assessed comprehension than global judgements of comprehension. As previously discussed (see Chapter 2), while the estimates obtained in Study 3 should not be directly interpreted as a measure of an individual's capacity to evaluate the status of their understanding of information (Paulewicz et al., 2020), research has suggested that regression slopes are sensitive to differences in underlying discrimination accuracy (Rausch & Zehetleitner, 2017). Given the measurement and analytic approach taken in Study 3, therefore, the observed results are arguably not purely an artefact of measurement and may provide some insight into underlying metacognitive processes associated with judgements of perceived comprehension.

Prevailing accounts of the weakly positive average association between measures of perceived and assessed comprehension suggest that individuals use various sources of information to infer their level of understanding (Koriat, 1997; Thiede et al., 2019). However, individuals typically use cues which are poorly diagnostic of their level of comprehension (Thiede et al., 2010). Due to a failure to make use of valid cues, or an unavailability of such cues, the internal metacognitive processes yield an evaluation which provides limited concordance with the true state of an individual's comprehension (Thiede et al., 2010; Wiley et al., 2016). While factors which may shape the internal judgement process have received considerable attention, research has less frequently examined how the association between perceived and assessed comprehension may be shaped by the approach to measurement (Weaver, 1990; Wiley et al., 2005). Addressing limitations in previous research, this study sought to examine whether the association between judgements and performance may be undermined by measurement error, due to a mismatch between the scope of the metacomprehension judgement and the coverage of the measure of comprehension (Dunlosky, Rawson & McDonald, 2002; Dunlosky et al., 2005; Wiley et al., 2005). The results of Study 3 are consistent with the view that aligning judgement scope and assessment coverage produces a stronger relationship between perceived and assessed comprehension.

The findings of this study reinforce the importance of ensuring a high degree of correspondence between the information offered for metacognitive evaluation and that which is assessed in the comprehension task in metacomprehension research. As highlighted by others, the nature of text comprehension presents a range of difficulties in constructing a valid and reliable measure of task performance (Weaver, 1990; Wiley et al., 2005). Researchers have further contended that complete overlap with the information offered for judgement is a minimal requirement for a valid measure of comprehension which does not constrain the estimated association with metacomprehension judgements (Thiede et al., 2009). However, obtaining complete correspondence between the coverage of judgements and performance measures presents a material challenge in metacomprehension research (Griffin, Mielicki & Wiley, 2019). As an alternative to increasing the coverage of the comprehension task, this study demonstrates that reducing the scope of the judgement prompt provides an effective method to increase the informativeness of metacomprehension judgements.

While a reduction in measurement error is a cogent explanation for greater predictive accuracy, factors relating to the internal judgement process may provide an alternative account for the observed results. Researchers have previously argued that global metacomprehension judgements may be challenging for individuals to provide, requiring the sampling or averaging of cues, or leading to the use of non-reading related cues, in order to form a judgement (Dunlosky & Lipko, 2007; Händel et al., 2020). As a result, the internal metacognitive processes produce evaluations of underlying comprehension which are more likely to be subject to inaccuracies. In contrast, judgements made of specific details within texts may be comparably easier and require no such sampling or averaging, increasing the predictive accuracy of metacognitive evaluations. The internal process of providing a metacognitive judgement, therefore, may be influenced by the volume of information offered for evaluation, with higher predictive accuracy more likely to be observed when judgements concern limited quantities of information (Dunlosky, Rawson & McDonald, 2002).

A potential mechanism through which judging a reduced quantity of information may both lower the difficulty and increase the predictive accuracy of the judgement is the use of the prompt-generated cues. Dunlosky, Rawson & McDonald (2002) have suggested that prompts targeting specific information may encourage participants to make focused retrieval attempts. Metacognitive experiences generated through recall attempts are then available to inform judgements of comprehension which directly correspond to the comprehension task, thereby increasing predictive accuracy (Koriat, 1997; Morris, 1990). However, this account may not fully explain the greater predictive accuracy observed here. Previous research suggests that individuals may not spontaneously make retrieval attempts, despite targeted judgement prompts, and are not able to independently verify the contents of their retrieval (Dunlosky et al., 2005; 2011; Dunlosky & Lipko, 2007). Moreover, even when judgement prompts are targeted to the level of single words, previous research has indicated that a delay prior to eliciting judgement appears instrumental in increasing predictive accuracy (Dunlosky et al., 2005; Rhodes & Tauber, 2011).

Nevertheless, metacognitive judgement processes may still have been influenced by cues generated by judgement prompts in this study. Local judgements may have been influenced by the perceived capacity to provide a response to the judgement prompt, without engaging in a recall attempt (Dunlosky, Rawson & McDonald, 2002; Dunlosky et al., 2005). Alternatively, processing fluency experienced in reading the judgement prompt may have informed local judgements (Koriat et al., 2004). Similarly, global judgements may have been influenced by metacognitive experiences related to processing, or answering, the local judgement prompts. Considering the distributions of ratings observed in Study 2 (see Appendix L and Figure 5.3) and Study 3, there was a considerable reduction in the highest ratings of perceived comprehension when text-level judgements were preceded by local judgements. This observation accords with research indicating global judgements are readily influenced by various salient features of the judgement environment (Griffin, Wiley & Thiede, 2019; Linderholm et al., 2008).

While the exact mechanism through which the association between perceived and assessed comprehension may increase remains unclear, two important observations arise from this study. Firstly, while reducing the coverage of the prompt drives an increase in the predictive relationship, local judgements do not strongly predict comprehension outcomes. If local judgements are more predictive due to greater alignment between the judgement scope and assessment coverage, this indicates that the weakly positive association observed is likely not solely attributable to measurement error in the comprehension task. Secondly, individual variability in the estimated predictive relationships was limited, with the predictive accuracy of all participants estimated to increase for local judgements and no negative associations estimated for either global or local judgements. This observation is not well accounted for in dominant theoretical accounts of metacomprehension judgements, which suggest that individual engage in various routes to judgement and rely on a range of cues (Griffin et al., 2009; Wiley et al., 2016). As cues vary considerably in their correspondence with actual comprehension, marked variability in estimated metacomprehension accuracy is observed (Chiang et al., 2010; Jee et al., 2006). In contrast, this study suggests that individuals took a similar approach to providing metacomprehension judgements and relied on cues which varied little in their diagnostic validity.

The anchoring and adjustment framework of metacomprehension judgements (Linderholm et al., 2008; Zhao & Linderholm, 2008), however, may provide a possible explanation of the observations described above. This framework posits that individual experiences or aspects of the judgement environment form the basis of metacognitive judgements, referred to as an 'anchor' (Scheck & Nelson, 2005; Tversky & Kahneman, 1974). Adjustments to judgements anchors are made using available cues, such as ease of retrieval or test feedback (Kubik et al., 2022; Zhao, 2022). However, due to insufficient adjustments, judgements remain biased toward the anchor (Tversky & Kahneman, 1974). The limited variability between individuals observed in Study 3 could result from minimal adjustments being made to the anchor points for global and local metacomprehension judgements, with participants utilising a different anchor point for each judgement type. This would suggest that anchor selection flexibly responds to judgement conditions and, simultaneously, is consistent across individuals. Alternatively, both judgement types may have relied on the same anchor, yet adjustments were more effective for local judgements. Unfortunately, these competing explanations cannot be evaluated using the current data. Further research is required, therefore, to explore whether homogeneity in anchor selection or differences in adjustment process might account for differences in metacomprehension accuracy.

5.4.3 Limitations

By reducing the scope of the judgement in Study 3, error in the measure of text comprehension introduced through assessing a sample of information from the text was reduced. However, another source of error in the measure of assessed comprehension may have contributed to variability in the relationship between perceived and assessed comprehension. In designing the local judgement prompts for this study, the wording of prompts was carefully controlled to avoid stating the correct answer or invalidating distractor option, while unambiguously prompting participants to evaluate their understanding of the same information later tested in the comprehension task. Accommodating both aspects within local judgement prompts was challenging and compromise was required for some prompts. An enduring concern, therefore, is that local judgement prompts may not have reliably directed participants to evaluate precisely the information targeted in the comprehension task, thereby introducing error. Consequently, the estimated predictive relationship between local judgements and assessed comprehension observed in this study may be an underestimate.

However, increased prompt specificity may influence measures of assessed comprehension, and, consequently, the association between perceived and assessed comprehension. Given these challenges, future research should consider how best to resolve this or explore methods to measure the impact on the findings.

A second limitation of this study is the observed question-level variability in the probability of observing a correct response on the comprehension task. Consistent with both Study 1 and 2, questions were estimated to contribute greater variance in observed comprehension outcomes than individual differences. While comprehension questions attempt to capture individual differences in understanding, driven by variability in the skills and processes required to comprehend text, attributes of questions can exert a considerable influence on responses. For multiple-choice questions, this includes the complexity of the question stem and response options, and the plausibility of distractor options (Ebel & Frisbie, 1991; Kayarkaya & Unaldi, 2020; Oakhill et al., 2014). To reduce variability related to question construction, prior to use in this study, the comprehension questions were reviewed using data obtained in Study 2. Alterations were made to questions with low discrimination and distractors which appeared problematic for high-performing participants, potentially translating into the slight reduction in the estimated question-level variability in Study 3. However, questions with considerable bias in response selection were not identified by this approach. For example, on the third question from the seventh text, over 70% of participants incorrectly responded that dengue virus is caused by mosquitos rather than a virus (see Figure 4.9 and Figure 5.8). While high-performing participants did not appear biased toward this distractor, the validity of assessed comprehension across participants may be affected. For lower performing participants, highly relevant distractors may have circumvented effortful consideration of correct responses (Ozuru et al., 2013). Since such measurement-induced variance can undermine the predictive relationship between perceived and assessed comprehension (Wiley et al., 2005), the findings reported here would benefit from replication with open-ended questions.

More generally, while judgements made at the level of specific details within a text are likely to be more informative of comprehension than judgements made at the level of a text overall, this finding does not tell us about the extent to which local judgements relate to comprehension of information not assessed in this study. Specifically, the relationship between local judgements and the comprehension of information not targeted by the local judgement prompt has not been evaluated. Local judgements of comprehension may be less predictive of comprehension outcomes for information not targeted by the prompt, given the lack of overlap between the scope of the judgement and coverage of the assessment measure (Dunlosky, Rawson & McDonald, 2002; Dunlosky et al., 2005; Wiley et al., 2005). However, since individuals rely on the same set of comprehension-building skills when reading a text, the understanding of various pieces of information within a text can be expected to associate at the individual-level (Eason et al., 2012). It is likely, therefore, that local judgements of comprehension are somewhat predictive of comprehension outcomes of information not evaluated in the judgement. For example, a local judgement evaluating information which requires an effortful bridging inference to comprehend may also be predictive of information within a text which requires the same level of processing. However, if there is little similarity between these processing requirements, it is likely that the predictive relationship between assessed comprehension and local judgements would reduce. Consequently, contrasting the relative informativeness of global and local metacomprehension judgements should be considered with respect to which comprehension outcomes are of interest.

5.4.4 Conclusion

Study 3 found evidence which indicated that metacomprehension judgements made in respect of specific details within a text are more informative of comprehension outcomes than judgements made at the level of the whole text. While the magnitude of the predictive relationship, between perceived and assessed comprehension, was greater for local comprehension judgements, both global and local judgements were weakly positively associated with comprehension outcomes. In addition, limited individual variability was observed in this relationship for both global and local judgements. Targeting judgements of perceived comprehension to specific information within texts can, therefore, be expected to increase the predictive validity of metacomprehension judgements across individuals. This provides a readily implementable method to increase the informativeness of reader panellists' judgements concerning the likely comprehensibility of particular details with health-related information. Nevertheless, while eliciting evaluations of specific details within texts yields a judgement which is more informative of the likely comprehensibility of that information, such judgements are not an equivalent substitute for formally assessing understanding.

To the extent that the estimated relationships inform of internal processes underlying metacomprehension judgements, the results of this study are consistent with the view that internally accurate metacognitive evaluations may be undermined by measurement error in the comprehension task. Alternatively, the use of cues which are more predictive of comprehension in response to local judgement prompts can account for the findings of Study 3. Regardless, metacomprehension judgements which concern ideas within texts remained only weakly predictive of comprehension outcomes across participants. This indicates that sources of error in this estimated relationship persists similarly across individuals. Future research should explore whether local metacomprehension judgements provide, at best, a noisy measure of latent understanding due to a shared tendency to base judgements primarily on a cue with limited predictivity, and the extent to which individuals can be oriented to evaluate an alternative primary judgement cue. Concomitantly, exploring methods to further reduce error in the measure of text comprehension would yield data against which theoretical accounts may be better evaluated.

6. General Discussion

The primary goal of the three studies reported here was to investigate whether metacomprehension judgements are likely to be informative of comprehension of health-related information. Individual differences in reading skill, background knowledge and standards of coherence were examined as possible sources of variability in the relationship between metacomprehension judgements and assessed comprehension. Further, whether metacomprehension judgements might be used more effectively was explored, through the consideration of semantic centrality and the scope of the judgement prompt. The capacity to achieve accuracy and precision in estimation, yielding data capable of robustly addressing the research questions, shaped decisions concerning the number of participants, stimulus texts, and comprehension questions per text included within each study. Analyses were informed by a consideration of the appropriateness and limitations of various statistical approaches. Sensitivity analyses were also conducted within each study to evaluate the robustness of the results to alternative methodological and analytical choices.

Shifting perspective to present an active reflection of the ideas and choices presented in this thesis, in this chapter I first discuss the literature motivating each of the research questions and how each study addressed these. Following this, I discuss the theoretical and practical implications of the findings, alongside the recommendations which arise from these results. General limitations shared across the studies are then explored, followed by the potential directions for future research. Finally, I conclude this chapter with a short reflection on the critical points which can be surmised from this research. Following this, a final chapter is presented which focuses on the guidance which may be offered to those making use of reader panels, reflecting the applied context which motivated the research reported within this thesis.

6.1 Research Questions and Main Findings

Previous research has frequently indicated that there is only a weakly positive association between judgements of text comprehension and assessed comprehension, on average (Dunlosky & Lipko, 2007; Lin & Zabrucky, 1998; Maki, 1998; Prinz et al., 2020a; Yang et al., 2022). Importantly, this association has been reported to vary considerably across the population, with some individuals showing near perfect accuracy in their metacomprehension judgements while others show a negative association (Chiang et al., 2010; Glenberg & Epstein, 1985; Jee et al., 2006). As described in Chapter 1, a range of factors have been investigated as potential moderators of metacomprehension accuracy, relating to characteristics of the individual, the text and the metacomprehension task (Lin & Zabrucky, 1998; Prinz et al., 2020a). Despite the volume of existent literature, the inconsistency amongst findings presented a challenge in determining the likely correspondence between reader panellists' judgements of health-related information and actual understanding. More critically, however, as discussed in Chapter 2, limitations in the statistical analyses and sample sizes, systemic within this area of research, rendered it difficult to robustly answer questions concerning the metacomprehension accuracy of reader panellists. These limitations motivated RQ1 and RQ2 (Chapter 3; section 3.1.1).

Study 1 was designed to address RQ1 and RQ2 by: i) providing an estimate for the average association between perceived and assessed comprehension, ii) quantifying the individual variability in this association, and iii) evaluating whether differences in reading ability or background knowledge may account for such individual variability. Consistent with previous research, the association between judgements of comprehension and performance on the metacomprehension task was found to be weakly positive. However, in contrast to previous observations and expectations, individuals were found to vary minimally from the average estimated association. In addition, individual differences in reading ability and background knowledge did not reliably moderate the association. Together, these results suggested a degree of commonality across individuals in the process of providing metacomprehension judgements.

The results of Study 1 were inconsistent with theoretical accounts which propose that individuals variously select sources of information, which differ widely in diagnostic validity, in order to infer their comprehension (Thiede et al., 2019). Therefore, I considered alternative perspectives which assume a shared reliance on a primary source of information. Identifying the shared cue for metacomprehension judgements, should one exist, would provide insight into what reader panellists' judgements principally convey. As discussed in Chapter 3 (section 3.4.2), however, previously suggested primary cues provided a limited account of the findings of Study 3 (Dunlosky, Rawson & Hacker, 2002; Griffin, Wiley & Thiede, 2019; Linderholm et al., 2008).

Drawing on Dunlosky, Rawson & Hacker's (2002) levels-of-disruption hypothesis and Kintsch's (1988) work on hierarchical semantic structures of text, I proposed that the extent to which a reader successfully establishes a coherent macropropositional structure of the text could plausibly be the primary cue (Chapter 4; section 4.1.1). Further, given the roles of reading ability, background knowledge, and a reader's standards of coherence in constructing a macropropositional structure, individual differences in these variables may influence metacomprehension accuracy. This proposed account motivated RQ3 and RQ4 (Chapter 4; section 4.1.2).

Study 2 was designed to address RQ3 by evaluating whether metacomprehension judgements showed a stronger association with information which featured in the macropropositional structure of the text (semantically central) than with information which did not (semantically peripheral). If macropropositional coherence served as the primary cue, it followed that judgements should be preferentially predictive of comprehension of semantically central information (see Chapter 4; section 4.1.1). In addition, to address RQ4, measures of reading ability, background knowledge and standards of coherence were included to estimate the influence of these variables on the association.

Contrary to expectations, the association between judgements and assessed comprehension was not found to vary with semantic centrality in Study 2. In addition, individuals were estimated to deviate minimally from the average estimated association between perceived and assessed comprehension. Further, limited evidence was observed in support of an effect of reading ability, background knowledge or standards of coherence on metacomprehension accuracy. Overall, the results of Study 2 indicated that macropropositional coherence was unlikely to serve as the primary judgement.

In evaluating the findings of Study 2, I considered that the level of association observed in Study 1 and 2 may be constrained by factors which are external to the judgement process, such as limitations in the measure of assessed comprehension (Wiley et al., 2005). Researchers have argued that a valid measure of comprehension must evaluate understanding to a sufficient breadth and depth (Griffin, Wiley & Thiede, 2019; Wiley et al., 2005; 2016) and that the quantity of text evaluated in the metacomprehension judgement must be captured in the comprehension task (Dunlosky, Rawson & McDonald, 2002; Lin & Zabrucky, 1998; Weaver, 1990). Failing to exhaustively assess whole-text comprehension, therefore, introduces error in the measure of comprehension, undermining the association with judgements (Griffin, Mielicki & Wiley, 2019; Weaver, 1990).

As discussed in Chapter 5 (section 5.1.1), consistent with previous suggestions (Dunlosky, Rawson & McDonald, 2002; Weaver, 1990), an increase in the association between perceived and assessed comprehension may be produced by aligning the quantity of text which is judged and assessed. This could be achieved by targeting a smaller quantity of text in the metacomprehension judgement. However, research indicates that reducing the judgement scope does not reliably increase metacomprehension accuracy (Dunlosky, Rawson & McDonald, 2002; Dunlosky et al., 2005; Lefèvre and Lories, 2004; Vössing & Stamov-Roßnagel, 2016). Further, these mixed results have been observed under conditions which are dissimilar to those experienced by reader panellists. It remained unclear, therefore, whether the weakly-positive association observed in Study 1 and 2 may arise due to measurement error issues. This lack of clarity, concerning the impact of scope of the judgement prompt, motivated RQ5 (Chapter 5, section 5.1.2).

Study 3 was designed to address RQ5 in two ways: i) comparing the attribution of response accuracy variance to judgements made at the level of the text and the level of an idea within a text and ii) calculating the difference in accuracy of predicted comprehension outcomes for judgements made at the level of the text and the level of an idea within a text. It was found that judgements concerning specific details within a text accounted for a greater proportion of the variability in comprehension outcomes than judgements concerning the whole text. However, only partial support was found for a difference in the accuracy of predicted comprehension outcomes between these judgement types. On balance, the findings suggested that reducing the scope of metacomprehension judgements to specific ideas within a text is likely to increase the association with assessed comprehension. Despite this, the magnitude of the increase in the association between perceived and assessed comprehension was estimated to be limited. This indicated that measurement error in comprehension task likely did not account for the weak association alone.

6.2 Implications and Recommendations

6.2.1 Theoretical

As discussed in Chapter 2 (section 2.1 and 2.1.1), several factors can lead to inaccuracies in the measurement of metacomprehension, such as judgement bias, performance ability, guessing, and the statistical properties of the estimated measures

(Benjamin & Diaz, 2008; Hattie, 2013; Higham & Higham, 2019; Masson & Rotello, 2009; Nelson, 1984; Rausch & Zehetleitner, 2017; Vuorre & Metcalfe, 2022). For these reasons, alongside concerns surrounding the capacity to identify and control confounding variables more generally, researchers have cautioned against interpreting estimates of metacomprehension as accurate measures of individuals' metacomprehension ability (Paulewicz et al., 2020). It is reasonable, therefore, to consider the extent to which my findings are informative of individuals' internal capacity to accurately evaluate their comprehension.

Given the motivation to evaluate the predictive relationship between measures of perceived and assessed comprehension on health-related texts, regression models were used in the present research. This approach to estimation and analysis also addressed shortcomings in alternative measures of metacomprehension, including the potential low reliability of difference scores, the loss of information in two-step analytic approaches, the need to incorporate hierarchical variance, and the capacity to explore different statistical assumptions (for a fuller discussion, see Chapter 2; section 2.1 and 2.2). However, regression slopes, as a measure of metacomprehension, are similarly vulnerable to non-metacognitive influences, potentially leading to inaccurate estimates of metacognitive ability (Rausch & Zehetleitner, 2017). Given this vulnerability, it was considered that individual-level estimated regression slopes would not be used as the basis for drawing strong inferences concerning individuals' underlying ability to evaluate their understanding of health information. Nevertheless, since regression slopes have been shown to capture differences in individuals' underlying ability to discriminate between stimuli states (Rausch & Zehetleitner, 2017), I considered that the estimated relationships reasonably permit speculation concerning internal metacomprehension processes.

Across each of the studies, two findings consistently emerged: i) the strength of the association between perceived and assessed comprehension was small and ii) limited variability was estimated across individuals. Assuming that my findings provide insight into underlying metacognitive processes, these observations indicate that individuals show fundamental similarity in their metacognitive evaluation and response processes, producing judgements which are comparably limited in their association with text comprehension. There are two major implications which arise from this, concerning how internal metacomprehension judgements may be formed and how the approach to experimental observation impacts on our conceptions of these internal processes.

Firstly, in contrast to the reported substantial individual-level heterogeneity in measures of metacomprehension accuracy previously observed (Chiang et al., 2010; Jee et al., 2006; Thiede et al., 2010; Griffin et al 2008), no evidence of high variability was found in any of the three studies conducted. The present findings are poorly accommodated within cue-based theoretical accounts which posit that individuals adopt distinctly different routes to judgement or variously select from cues which are highly variable in their diagnostic validity (Griffin et al., 2009; Koriat, 1997; Thiede et al., 2010). Instead, the findings are more plausibly accounted for by accounts which propose a tendency for individuals to predominantly attend to a single judgement cue.

Previous suggestions of a single, default judgement cue, however, do not provide a full explanation of the metacomprehension behaviours I observed. For example, given the proposal of a default cue of an expectation for a memory-based test, it is unclear what default cue could be used in the absence of informing readers of an upcoming test (Griffin, Wiley & Thiede, 2019). Similarly, the impact of the severity, or the successful resolution, of disruptions to text comprehension processes (Baker, 1979) are underspecified within accounts which suggest that disruptions to comprehension serve as the default judgement cue

(Dunlosky, Rawson & Hacker, 2002). Further, my observations are inconsistent with accounts which suggest that the total amount of stimuli-related information which is directly accessible in memory serves as the primary judgement cue (Dunlosky et al., 2005; Koriat, 1993, 1995). Eliciting local judgements in Study 3 appeared to reduce the magnitude of global judgement ratings, compared to Study 2, despite local judgements likely increasing the ability to recall information from the text.

While a default judgement cue would account for homogeneity in metacomprehension accuracy, my findings indicate that a suitable account of judgement processes must be more nuanced than the assumption that all individuals simply attend to a single judgement cue. Though not at the levels reported in previous research, some individual variability in the relationship between perceived and assessed comprehension was observed. In addition, metacomprehension judgements were influenced by the scope of the judgement prompt. In Study 3, judgements concerning comprehension of specific details were found to be simultaneously lower in magnitude and more strongly predictive of assessed comprehension. Further, eliciting judgements concerning specific ideas prior to judgements of the whole text appeared to reduce the magnitude of the latter, relative to judgement magnitudes observed in Study 2.

Arguably, my findings indicate that metacomprehension judgements are characterised by two features. Firstly, there is a core cue which is combined with other sources of variance. This assumption would produce the observed similarity across individuals in the overall magnitude of the association between perceived and assessed comprehension, while accounting for the limited differences between individuals. Secondly, the selection of judgement cues is influenced by external factors. This assumption would produce the differences observed between the judgements concerning the whole-text and specific ideas within texts. Further, this assumption would account for the observed impact of eliciting judgements concerning specific ideas prior to eliciting whole-text judgements. These proposed features are consistent with an anchoring and adjustment framework of metacognitive judgements (Linderholm et al., 2008; Tversky & Kahneman, 1974).

Discussed in Chapter 1 (section 1.3.1), the anchoring and adjustment account of metacognitive judgements posits that a single source of information forms the basis of metacognitive judgements, from which individuals then adjust to arrive at a judgement (Linderholm et al., 2008; Scheck & Nelson, 2005; Tversky & Kahneman, 1974; Zhao & Linderholm, 2008). According to the anchoring and adjustment account of metacomprehension judgements, judgement magnitudes should change when uncertainty in the judgement is reduced or when individuals are encouraged to attend to experience-based cues (Zhao & Linderholm, 2008). Support for this can be seen in Griffin, Wiley and Thiede's (2019) research, in which informing participants about the nature of the task was found to reduce judgement magnitude in the first experiment. In addition, effects on judgement magnitudes have been observed when individuals engage in tasks which purport to encourage individuals to attend to experience-based cues (Prinz et al., 2020b; Wiley et al., 2016), such as summarising the text (Anderson & Thiede, 2008; Thiede et al., 2010). However, several studies have failed to demonstrate such effects. Judgement magnitudes have been observed to remain stable regardless of whether or not participants are informed about the nature of the comprehension task (Griffin, Wiley & Thiede, 2019; Wiley et al., 2016), when participants are asked to summarise the text (Madison & Fulton, 2022), or to engage in other situation-model based interventions, such as self-explanation during reading (Griffin et al.,2008; Rawson et al., 2000).

With respect to my findings, the anchoring and adjustment account of metacomprehension judgements (Zhao & Linderholm, 2008) arguably provides an incomplete account. The use of a default cue which is shared across individuals, as indicated by my findings, is not stipulated by this account. Instead, the anchoring and adjustment account suggests only that individuals may typically use an anchor based on performance expectations (Linderholm et al., 2008; Zhao & Linderholm, 2008). In addition, while the effect of judgement prompt observed in Study 3 may be attributed to reduced task uncertainty which leads to less anchoring on knowledge-based cues, this theoretical account does not describe what source(s) of information individuals use instead

More generally, the anchoring and adjustment account of metacomprehension judgements does not describe in detail the process of adjusting or the factors which govern the effectiveness of this process. Empirical research indicates that individuals appear to automatically adjust their judgements using experience-based cues when not explicitly encouraged to do so (e.g., Rawson & Dunlosky, 2002), yet fail to adjust judgements when provided with more precise information about the comprehension task (e.g., Dunlosky et al., 2005; Wiley et al., 2016). As a consequence of the limited specification of anchoring and adjustment processes, it is challenging to discern whether such observations can be accommodated within, or are in fact incompatible with, this account. Similarly, it is unclear whether observed changes in judgement magnitudes correspond to changes in the underlying judgement anchor or in the use of adjustment processes.

How, then, might we satisfactorily account for the present findings and those within the wider literature? Previous research suggests that a range of factors can influence metacomprehension accuracy, as individuals readily adjust their metacomprehension judgements in response to features of the task (Prinz et al., 2020a). According to cue-based accounts, failing to observe an effect of a manipulation on metacomprehension accuracy occurs when the manipulation fails to prompt individuals to utilise cues which are highly diagnostic of their understanding of information which is assessed on the comprehension task (Dunlosky & Lipko, 2007; Griffin, Mielicki & Wiley, 2019; Koriat, 1997). Explanations for why individuals fail to base metacomprehension judgements on valid cues principally focus on the lack of their availability, potentially caused by limited resources available to monitor for valid cues during reading (Dunlosky & Rawson, 2005; Griffin et al. 2008; Millis et al., 1998).

Less well empirically evaluated, due to the difficulty in accurately verifying actual cue usage, is what drives the choice of judgement cue(s) from those available. Despite this challenge, researchers have suggested that cue use may be driven by salience, heuristics or knowledge-driven processes (Dunlosky, Rawson & Hacker, 2002; Koriat, 1997; Griffin et al., 2009; Griffin, Mielicki & Wiley, 2019; Linderholm et al., 2008; Thiede et al., 2010; Wiley et al., 2016; Zhao & Linderholm, 2008). In response to task manipulations in which participants are not explicitly instructed to form their comprehension judgements on the basis of the putative cue, salience-driven cue selection arguably more readily accounts for observed differences in metacomprehension accuracy. For example, considering the observed impact on metacomprehension accuracy, it may be reasonable to consider that engaging in situation-model based interventions increases the salience of highly predictive cues concerning the situation-model level of understanding (Prinz et al., 2020b). Similarly, manipulating the reading ease of stimulus texts may increase the salience of cues concerning perceived reading fluency (Dunlosky et al., 2006; Maki et al., 1990; Rawson & Dunlosky, 2002).

To propose a potential account for my findings, salience-driven cue selection may be productively combined with concepts from the anchoring and adjustment framework. Firstly, individuals take the most salient cue as a judgement anchor. In the absence of task manipulations, this may be processing fluency or the quantity of information available to recall (Dunlosky et al., 2005; Koriat, 1993; Koriat et al., 2004; Morris, 1990). If an alternative cue is made salient, such as disruptions to comprehension, test expectations, or reading fluency, individuals will readily utilise this as an anchor (Dunlosky, Rawson &

Hacker, 2002; Griffin, Wiley & Thiede, 2019; Rawson & Dunlosky, 2002). Secondly, limited adjustments to the anchor are made on the basis of available knowledge- and experience-based cues (Epley & Gilovich, 2006). Given this mixture of bottom-up salience-driven anchor selection and top-down adjustment process, we may reasonably expect to see relative homogeneity in metacomprehension behaviours when the characteristics of the task are held constant across individuals. Further work is required, however, to describe adjustment processes more fully. In addition, the assumed homogeneity in metacomprehension behaviours in response to experimental conditions is based on the trends I observed and, therefore, requires independent replication.

The need to further evidence the limited variability I observed in the relationship between perceived and assessed comprehension also relates to the second major implication of my findings. The measurement approach adopted in the three studies was driven by a consideration of the most appropriate method to quantify the predictive association between measures of perceived and assessed comprehension, given the goals of the present research and experimental design. To address the issues identified in Chapter 2 (section 2.1 and 2.2) generalised multilevel regression models were considered most appropriate (Baayen et al., 2008; Gelman & Hill, 2007). This approach allowed for the capacity to informatively estimate the uncertainty in multiple parameters, without the violation of statistical assumptions or loss of information. However, this method differed substantially from those typically employed in metacomprehension research (Griffin, Mielicki & Wiley, 2019; Prinz et al., 2020a). This difference in analytic approach likely accounts for the contradictory findings regarding individual-level variability observed here and elsewhere.

As discussed in Chapter 2 (section 2.3) and in the discussion of findings within each study, it is likely that analytic approaches typically used within metacomprehension research approaches do not provide accurate and precise individual-level estimates. This is primarily due to an insufficient quantity of data, as metacomprehension studies, on average, estimate measures of metacomprehension using five pairs of judgement and response scores (Prinz et al., 2020a). As noted by Griffin, Mielicki and Wiley (2019, p. 629), "it is easy to find fault with any correlational method based on only six judgment–test pairs per participant". Indeed, given a true Pearson's correlation of approximately 0.25, stability in accuracy of the estimate requires over 150 observations (Schönbrodt & Perugini, 2013), while a precision level of 0.1 in the 95% confidence interval for the estimate requires over 300 observations (Harrell & Slaughter, 2020). Given that obtaining this volume of data is unachievable in metacomprehension research, interpreting individual-level estimates should be strongly advised against.

Despite the inaccuracy and imprecision of measures of metacomprehension accuracy estimated at the individual-level, aggregating estimates across both studies and individuals may provide a more accurate measure of the average level of metacomprehension ability. Conditional on individual-level estimates of metacomprehension accuracy being independent and identically distributed (as approximately observed here), the magnitude of cross-study-level estimates of metacomprehension accuracy will be located closer to the true average value of metacomprehension accuracy and will have lower variability (see Appendix P for a full account and example of this). If these assumptions were true, we would expect to observe relative stability in cross-study analyses of study-level averages of metacomprehension accuracy. This is, in fact, what is reported in the literature, with cross-study analyses indicating an average correlation between perceived and assessed comprehension of between 0.18 to 0.27 (Dunlosky & Lipko, 2007; Lin & Zabrucky, 1998; Maki, 1998; Prinz et al., 2020a; Yang et al., 2022). This explanation may account for why the present research successfully replicated the weakly-positive average association but failed to demonstrate high levels of individual variability.

There are clear implications of the approach to measurement on theoretical accounts of metacomprehension judgements, given the above. Correctly identifying the characteristics of the phenomenon we wish to describe is crucial in building theories which can express behavioural processes with high fidelity. It is essential, therefore, that our methods of measurement and analysis do not introduce artifacts which lead to spurious conclusions. As described above, whether individuals are highly variable or are markedly similar in their metacomprehension behaviours will influence our conceptions of the internal judgement process and of which theoretical accounts are plausible.

Future metrics of metacomprehension ability may prove more informative than the analytic approach adopted here, such as the extension of signal-detection theoretic based-measures to the experimental designs typical in this area of research (Fleming, 2017). More complex mathematical models which define the process through which an internal judgement state is translated into a discrete judgement and which represent underlying metacomprehension ability as a model parameter would likely be particularly insightful. Such models would permit the direct estimation of a measure of metacomprehension which arises from a well-defined process, making underlying aspects of metacomprehension judgements more amenable to experimental investigation. Until such time, as demonstrated in the present research, the tools to improve practice are already available. Theoretical development within the field of metacomprehension research would benefit from the use of analyses which better allow us to appropriately handle and quantify the uncertainty in our observations.

6.2.2 Practical

Providing high quality, written information for patients plays an important role in healthcare, allowing patients to make informed decisions, facilitating engagement, and improving outcomes (Department of Health, 2003; Audit Commission, 1993). To this end, reader panels are employed to provide a valuable, non-clinical evaluation of the likely comprehensibility of draft documentation (e.g., Cambridge University Hospitals NHS Foundation Trust, 2023). However, the present findings highlight the need to carefully consider what can be interpreted from reader panellists' responses.

Based on the findings of Study 1, 2 and 3, it can be concluded that a low judgement of perceived comprehension indicates non-optimal understanding. For texts rated at the lowest level of perceived comprehensibility, a notable portion of information is likely to be poorly understood. Low judgements, therefore, signal that communicators should revise their documentation. In contrast, a high judgement of perceived comprehension indicates that a large amount of the text is likely to be understood. However, observing the highest levels of perceived comprehensibility does not indicate which aspects of the text are understood well and which are not, nor exactly how much information remains poorly comprehended. High endorsements of the comprehensibility of health information may, therefore, be misleading: judgements may indicate a high level of understanding when, from a clinical perspective, this may be deficient in important ways.

My findings raise questions concerning the sufficiency of current practice: what do health communicators want patients to understand from health-related documents and how does the review process evaluate this? Pragmatically, patients should have a sufficient breadth and depth of understanding which allows them to make informed decisions and correctly adhere to medical guidance (Department of Health, 2003). Adequate comprehension of certain information within health-related texts is, therefore, more important than other information. For example, document production guidelines state that, alongside providing instructions, an explanation for the reason for the instruction should be given (Department of Health, 2003). Comprehension of the explanation is arguably incidental to that of the instruction.

Unfortunately, the findings of Study 2 indicate that text-level judgements do not distinguish between either understanding of the main information an author wishes to communicate or comprehension of supplementary details. Further, while asking individuals about specific details within health-related texts appears to be an effective and easily implementable method to learn about the comprehension of particular details, these judgements remain limited. The findings of Study 3 suggest that a high judgement does not indicate that any specific piece of information is sufficiently well understood, nor does a low judgement guarantee that the opposite is true. For health communicators who are concerned about whether certain aspects of documents are likely to be understood, these judgements offer limited utility.

Ultimately, my findings indicate that judgements of perceived comprehension should not be considered a gold-standard approach for evaluating the likely comprehension outcomes of patients engaging with written health information. Across the studies, individual differences in reading ability and health-related background knowledge were more predictive of comprehension than judgements of comprehension. This suggests a more effective evaluation of patient information may be to simply consider whether the text is suitable given the distribution of reading-related skills within the document's target population (Department of Health, 2003).

More worryingly, however, across the studies, judgements concerning both the quality of the writing and whether the language was perceived to be patient friendly did not reliably associate with comprehension outcomes. Questions concerning writing quality and patient-friendly language feature frequently in reader panel evaluations (e.g., Chesterfield Royal Hospital NHS Foundation Trust, 2023; Mid Cheshire Hospitals NHS Foundation Trust, 2023). Communicators may believe that a document rated highly in these aspects is likely to engender high levels of text comprehension, yet I observed no evidence which supports this view. Evidently, in gauging patients' likely understanding of health information, perception-based measures provide limited, or even potentially misleading, feedback.

In what capacity, if any, should reader panels be used in the evaluation of patient information? It may be suggested that, despite the limitations of the current format of reader panel evaluations, obtaining negative feedback provides a useful indication that patients are likely to experience comprehension problems. When negative judgements are observed at the level of specific information within texts, these perceptions become more informative and can identify which elements of the document are likely associated with comprehension difficulties. Moreover, engagement with non-expert, patient populations provides valuable feedback relative to the wider social context within which health documents exist, such as whether the document is respectful of cultural differences (Department of Health, 2003). Arguably, however, the current usage of reader panels represents a missed opportunity to assess, with greater rigour, whether the core points of information an author wishes to communicate are likely to be understood by patients. Greater insight could be obtained by incorporating an objective measure of assessed comprehension. It is recognised, however, that this would entail increased input from both communicators and reader panellists. Further consideration is required to ensure that an objective measure could be suitably accommodated and would not dissuade reader panellist participation.

6.3 General Limitations

Each of the studies reported in this thesis are not without limitations, as described in sections 3.4.3, 4.4.3, and 5.4.3, in Chapter 3, 4, and 5, respectively. Here, I focus on a key limitation: the generalisability of the findings to the wider population of adult readers. Participants recruited in Study 1 and 2 responded to questions taken from the QRI (Leslie & Caldwell, 2017) to provide a measure reading ability. The selection of test material from the QRI was driven by an evaluation of the average level of reading ability in the UK adult population (Kuczera et al., 2016; Office for National Statistics, 2013). Based on this, it was considered that the average highest level of qualification may be between level 2 and level 3, equivalent to achieving five or more GCSEs or two or more A-levels. The passages selected in the QRI corresponded to upper-middle school, the approximate school age in which UK children would be studying for their GCSEs.

It was anticipated that performance on the QRI passages would roughly reflect the national average, with the majority of participants responding correctly to half or more of the of the QRI questions. Performance on the QRI generally supported this, with a median score of 70% observed both Study 1 and 2. Although the measure of reading ability was sufficient to capture variation in the sample of recruited participants, very few observations were collected from individuals with low levels of reading ability: only 12 and 29 participants in Study 1 and 2, respectively, scored below 50% on the task. As a consequence of the lack of data from participants with low reading ability, the findings reported may not generalise well to individuals who struggle with text comprehension. A sample which consists of a markedly greater proportion of low ability readers may lead to different conclusions, perhaps revealing lower metacomprehension accuracy (Griffin et al., 2008; Ozuru et al., 2012).

While low ability readers may be less likely to volunteer as reader panellists, the possibility that the results observed here may not accurately reflect the metacomprehension ability of low ability readers is concerning when considered in a wider context of understanding health information. As low levels of literacy are associated with poorer health (Weiss et al., 1992), in order to encourage positive health outcomes (Department of Health, 2003), it is imperative that efforts are made to try to improve understanding for these individuals. Attempts to identify a lack of understanding in patients with low literacy could be further frustrated by assuming that their metacomprehension judgements are predictive of understanding (Eason et al., 2013). The findings reported here, therefore, should be interpreted as indicative of metacomprehension accuracy in fluent adult readers.

More reassuringly, considering the observed variability in the measure of health-related background knowledge alongside the findings of previous research suggests that individuals with low health literacy may have been well-represented in the study samples. The HLVA was originally designed to provide a measure of health-related literacy in individuals with English as their first or second language (Ratajczak, 2020). Health literacy is a construct which is broadly defined as the capacity to comprehend and use written health-related information (Kickbush, 2001). Studies of health literacy have found that almost a quarter of participants are considered to have low health literacy (Lee et al., 2010). Worse still, at least a third of the population in the multiple European countries may have inadequate or problematic health literacy (Sørensen et al., 2015). Based on the findings of Ratajczak (2020) and Lee et al., (2010), a rough approximation suggests that a score of 7 or below on the revised HLVA may be considered indicative of low health literacy.

Consistent with estimated levels of low health literacy in the population, a threshold of ≤ 7 on the HLVA would correspond to observing low health literacy in approximately 25% and 28% of participants in Study 1 and 2, respectively. This suggests that individuals likely to have low levels of health literacy were reasonably well captured within the samples. The findings reported, therefore, may be expected generalise well across levels of health literacy. However, this calculation is based on responses collected on the original HLVA (featuring 22 items, open-ended verbal responses, and partial credit scoring). Evaluating the association between existing measures of health literacy and the revised HLVA used in the present studies would provide greater confidence in the generalisability of the findings.

6.4 Future Directions

Future experimental work is required to address the identified limitations in the present research. Principally, greater recruitment of individuals with low levels of reading ability would provide insight into the generalisability of these findings to non-fluent adult readers (Griffin et al., 2008). Further, the use of open-ended questions to measure assessed comprehension would quantify the extent to which the estimated association between perceived and assessed comprehension is undermined by non-relevant features of the comprehension task itself (Ebel & Frisbie, 1991; Kayarkaya & Unaldi, 2020; Oakhill et al., 2014; Ozuru et al., 2013). In addition, re-analysing previously collected study data using the analytic approach adopted here would clarify whether the observed homogeneity in the predictiveness of metacomprehension judgements between individuals results from merely differences in analytic choices. However, to meaningfully drive progress this field and achieve a greater understanding of metacomprehension judgements, what is arguably required is much increased theoretical specificity (Navarro, 2021).

Following decades of research, proposed theories of metacomprehension offer considerable levels of detail and yet can only make, at best, ordinal predictions (Dunlosky, Rawson & Hacker, 2002, Griffin, Mielicki & Wiley, 2019; Zhao & Linderholm, 2008). Further, the hundreds of published empirical studies provide a limited capacity to discriminate between various accounts of the metacomprehension processes, with findings remaining consistent with multiple competing theories (Yang et al., 2022). Within cue-based accounts, the lack of an observed effect in response to a task manipulation can be attributed to individuals failing to utilise alternative cues or a failure of the manipulation to make alternative cues available and salient (e.g., Dunlosky et al., 2005; Jaeger & Wiley, 2014). As such, since actual cue use cannot be verified, null results can be readily explained without undermining the theoretical account under investigation. Concomitantly, observing a significant difference in metacomprehension behaviours, however small or large, is considered consistent with the directional effect predicted by the supporting theoretical account (e.g., Prinz et al., 2020b). As a result, most empirical research conducted can be considered a weak test of theory, with the capacity to empirically falsify these theories potentially questionable (Meehl, 1978, 1990; Popper, 1963). Without specific predictions and the ability to observe concrete refutations of proposed theories, research cannot easily provide evidence which indicates a failure of a theoretical account (Meehl, 1990; Navarro, 2021; Popper, 1963).

Existing theories of metacomprehension judgements have proved fruitful in identifying the conditions which improve the observed correspondence between perceived and assessed comprehension (Prinz et al., 2020b). However, our understanding of what cues are available to individuals at any particular instance of judgement, how cue selection processes operate, how cues may be combined, and how an internal judgement state is translated into an overt response, remains impoverished. To address this, theoretical accounts are required which clearly and coherently formalise metacomprehension judgement processes in far greater detail. Through better-specified theories, highly specific and testable predictions can be generated, guiding empirical research which can rigorously evaluate support for a theory, in turn, progressing our understanding (Popper, 1963).

Mathematical models are arguably essential to quantify the constructs and their interrelations, in addition to explicitly expressing our assumptions (Navarro, 2021). Process models which characterise responses as the classification of an underlying evidence distribution into discrete outcomes, according to some threshold(s), (e.g., Maniscalco & Lau, 2012) provide a productive starting point for this work. Moreover, mathematically specifying a generative model permits the simulation of metacomprehension behaviours which could be evaluated against the volumes of existing empirical data. Despite the considerable challenge that achieving this presents, such work is fundamental in progressing accounts of metacomprehension judgements beyond descriptive adequacy.

6.5 Concluding Remarks

Across the three studies reported here, metacomprehension judgements weakly predicted an individual's comprehension of health-related text. This was observed for the comprehension of information more or less well-connected within the semantic hierarchy of the text. The association was similarly weak for judgements made in respect of both the whole text and specific details within texts, though the latter was slightly more predictive of comprehension. In samples of adult readers, varying in reading ability, background knowledge and standards of coherence, individual variation in the predictive association was limited, with these variables failing to reliably moderate the relationship. Overall, these findings suggest that reader panellists' judgements of perceived comprehension do correspond to underlying comprehension and, therefore, could serve as a valid tool in evaluating the comprehensibility of patient information. However, as a proxy for understanding which shows limited predictive correspondence with actual comprehension outcomes, judgements of perceived comprehension should not be considered a robust method for reviewing written health information. Practitioners who make use of these judgements should carefully consider the interpretation of responses and the potential consequences of obtaining inadequate information.

7. Insights for Those Making Use of Reader Panellists' Judgements in the Evaluation of Written Health Information

In this final chapter, the data collected in Study 2 and 3 are integrated to explore, more fully, the correspondence between subjective judgements of comprehension and understanding of health-related texts. Given the applied context of the present research project, the primary goal of this chapter is to provide guidance for those making use of reader panels in the review of written health information. In addition, recognising the contrasting approach taken within this research project and that adopted within metacomprehension research, key contextually driven differences in methodology and analysis are discussed, with the latter explored through a presentation of alternative analyses. This chapter is presented as a self-contained in the style of a journal article, with references to the work contained in previous chapters consequently limited.

7.1 Introduction

Written health information is an important component of patient healthcare. Providing high quality documents to patients can encourage engagement, facilitate adherence to guidance, and promote positive clinical outcomes (Audit Commission, 1993; Department of Health, 2003). In order to yield such benefits, it is essential that health-related texts are understandable to patients. To this end, such documents are produced according to sets of published guidelines (Department of Health, 2003). In addition, draft documentation typically undergoes review via reader panels prior to publication (e.g., Somerset NHS Clinical Commissioning Group, 2019). Reader panels consist of small groups of individuals who provide feedback on draft health documentation. The present research focuses on this latter element of health text production: considering the value of subjective evaluation in the creation of understandable written health information.

In evaluating draft health documentation, reader panellists provide subjective judgements concerning several aspects of the text. Reader panellists are typically asked to consider whether the document makes sense, how easily understandable the information is, whether the language is patient-friendly, and whether the writing of the text is of good quality (e.g., Cambridge University Hospitals NHS Foundation Trust, 2023; Mid and South Essex Integrated Care System, 2023). In obtaining and utilizing the information generated through the reader panel review process, it is implicitly assumed that reader panellists' evaluations provide an accurate measure of the comprehensibility and quality of draft documentation. This assumption underpins the view that reader panel evaluation offers an effective method to assess whether written health information is fit for purpose.

Despite the widespread use of reader panels, there is a surprising lack of research into the role of subjective evaluation in helping to produce understandable written health information. As a result, a number of questions remain regarding the use of reader panels, including whether reader panellists vary in their capability as reviewers and whether their judgements may be shaped by particular aspects of health-related texts. Most critically, however, it remains unclear whether reader panellists' subjective judgements of draft documentation provide a valid measure of the quality of a document. Consequently, therefore, health communicators cannot be certain that a judgement of high comprehensibility translates to a document which is likely to be understood.

Worryingly, previous research which has been conducted within an educational context indicates that subjective judgements of text comprehension are only weakly predictive of assessed understanding, on average (Prinz et al., 2020a; Yang et al., 2022). In these studies, students are asked to read excerpts from educational texts in advance of taking a test on the content. After reading each text, students make a judgement concerning their perceived level of learning or understanding gained from the text, referred to as a ‘metacomprehension judgement’. Numerous studies and recent meta-analyses have indicated that the average correlation between metacomprehension judgements and subsequent performance on a test of the material does not exceed 0.3 (Dunlosky & Lipko, 2007; Lin & Zabrucky, 1998; Maki, 1998; Prinz et al., 2020a; Yang et al., 2022). Furthermore, at an individual-level, there is marked variability in both the magnitude and direction of the association between metacomprehension judgements and assessed comprehension (Chiang et al., 2010; Glenberg & Epstein, 1985; Jee et al., 2006).

In accounting for the limited correspondence between metacomprehension judgements and test performance, it has been proposed that students rely on inferential judgement processes when making decisions about their learning (Griffin, Mielicki & Wiley, 2019; Koriat, 1997). In evaluating whether educational materials have been sufficiently comprehended or whether further study is required, self-regulated learning decisions are suggested to be guided by various judgement cues (Griffin et al., 2013). Judgement cues include knowledge or belief-based information, such as the amount of material to be learned, and experience-based information, such as the processing fluency experienced during learning (Griffin, Mielicki & Wiley, 2019). The accuracy of a metacomprehension judgement is determined by the extent to which such cues relate to a student’s understanding and memory of the material, with students typically selecting suboptimal judgement cues (Thiede et al., 2019).

Whilst the observed poor accuracy of metacomprehension judgements amongst students is highly concerning, these findings and accounts of students’ judgement processes may not generalize to the context of reader panellists’ evaluations of health-related texts. In studies which examine students’ judgements of educational materials, text comprehension is assessed based on the mental representation constructed during the reading of the text. The mental representation of the text, or ‘situation model’ (Kintsch, 1998), refers to the network of meanings and connections drawn between information extracted from the text and existing knowledge (Kintsch & Van Dijk, 1978; Trabasso et al., 1984; van den Broek, 1994). Essential to the measurement of understanding, to evaluate the accuracy of metacomprehension judgements, the quality of the mental representation held in memory is assessed, without further reference to the text (Wiley et al., 2005). Students' performance on the comprehension test is then operationalised as a sum score of the response accuracy across multiple test items, providing a holistic measure of overall performance, reflecting the quality of the text representation constructed during reading (Wiley et al., 2005). It is arguably the case that the behaviour captured using this approach diverges from that which is of most interest to those making use of reader panels.

Firstly, in the context of written health information, communicators are interested in a patient's capacity to both understand and to use such texts to make appropriate health decisions. Reflecting this joint concern, the term 'health literacy' is used to refer to an individual's ability to both comprehend and to use health-related texts (Institute of Medicine, 2004; Ratzan & Parker, 2000). In practice, therefore, it is assumed that such documents are available for patients to actively refer to when required. This use of texts markedly contrasts with that of educational materials, for which students are required to both understand and commit information to memory in order to perform effectively on assessments. There is no such expectation on patients to memorise health information. In investigating reader panellists' judgements, it is important to accommodate this difference in priorities within the experimental design, as assessing understanding in the absence of health-related texts would not reflect how these documents are used in-situ and would fail to appropriately explore the validity of reader panellists' judgements. In the present context, therefore, it is necessary to explore the extent to which subjective evaluations of health information are predictive of the capacity to use such documents to demonstrate comprehension.

Secondly, the functional role of health-related documents within patient healthcare shapes what is considered to constitute successful comprehension of the text. In metacomprehension research, a student's comprehension of written material is predominantly evaluated using a sum score of response accuracy across test items, capturing an overall measure of understanding across the text (Wiley et al., 2005). However, in a health context, a greater regard is often afforded to demonstrating understanding of elements within patient information rather than comprehending the document in its entirety. For example, communicators may be more concerned with ensuring a therapeutic treatment regimen is understood, rather than gauging overall understanding of a health document which can include information not personally relevant for a particular patient. Adequate comprehension of health-related text can, therefore, be understood in terms of successfully locating and understanding elements of the text which permit the reader to make informed decisions about their health (European Commission, 2009; Institute of Medicine, 2004). Measuring performance at the level of response accuracy to individual comprehension questions more closely aligns with this conceptualisation of understanding than calculating a sum score across items. In the present context, therefore, a fine-grained measure of comprehension is required to appropriately examine the relationship between reader panellists' subjective evaluations and the capacity to demonstrate understanding of health-related information.

The two critical contextual differences between educational and health-related texts, discussed above, give rise to two different methodological choices required to appropriately investigate subjective evaluations of text: (i) to assess understanding when the health-related text is available for reference rather than to assess the quality of the text representation held in memory, and (ii) to analyse item-level response accuracy rather than sum score performance. The necessary divergence in approaches both limits the extent to which previous findings can be leveraged to provide insight into the utility of health text evaluation via reader panels and indicates that existing theoretical frameworks are not appropriate to accommodate reader panellist's judgement processes. It remains unclear, therefore, whether eliciting subjective evaluations from reader panels provides a valid approach to gauging the quality of draft health documentation. To equip health communicators with a better understanding of reader panels, further research is required to investigate whether reader panellists' judgements provide a useful tool in producing understandable patient information.

7.1.1 The Present Research

What should reader panellists be asked to do in order for their input to be effective in helping to produce understandable patient information? For those making use of reader panels, how should panellists' judgements be utilised in evaluating draft health documentation? Given the absence of prior research which reflects the context in which reader panellists' judgements are made and used, it is challenging to answer these questions. To this end, secondary data will be used in the present research which permit an examination of the strength of the relationship between individuals' judgements of perceived understanding of health information and item-level analyses of responses to comprehension test questions administered with the text present.

In order to examine item-level of response accuracy, an analysis such as logistic regression is required. Given this analysis approach, the magnitude and sign of the estimated beta slope coefficient represents the strength and direction of the relationship between perceived understanding and performance on an individual comprehension question, on average. Extending such an analysis to a hierarchical framework additionally allows for simultaneous estimation of the average beta slope coefficient across participants and each participant's deviation from this average (Baayen et al., 2008; Gelman & Hill, 2007). Further, Bayesian estimation is argued to allow greater scope for expression of uncertainty within this framework (Kruschke, 2013). It can be argued, therefore, that a Bayesian multilevel logistic regression model is best suited to achieve the analytic aims of the present research.

Critically, however, such an analysis of item-level responses deviates from the traditional analytic approach within metacomprehension research. Typically, metacomprehension studies report calculating a participant-level correlation to provide a measure of an individual's ability to discriminate between levels of understanding or learning across a set of texts, often referred to as 'relative accuracy' (Bol & Hacker, 2012; Nelson, 1984). A t-test or ANOVA is then conducted to evaluate whether a statistically significant difference is observed in average the participant correlation between experimental conditions (e.g., Griffin, Wiley & Thiede, 2019). Given differences between this approach and multilevel logistic regression, in addition to the novelty of using the latter to evaluate the relationship between perceptual judgements of understanding and performance on a comprehension task, an analysis of participant-level correlations will also be conducted. The presentation of both analyses, using the same empirical data, will provide scope for a consideration of the conditions under which concordance or divergence of findings may be observed.

In addition to permitting both item-level and sum score analyses, the secondary data also allows for an exploration of possible sources of variability the relationship between judgements and understanding, thereby offering the potential for greater insights for those making use of reader panels. Discussed in greater detail below, two factors will be examined in the present research: (i) the impact of the scope of the prompt used to elicit perceptual judgments and (ii) the contribution to perceptual judgements of successfully understanding information at varying levels of importance in the semantic structure of the text.

Firstly, with respect to the scope of the judgement prompt, when reviewing draft documentation, reader panellists are asked to provide evaluations which pertain to the whole document. This approach to eliciting judgements is similar to that typically used in metacomprehension research, with judgement prompts predominantly inviting the participant to evaluate the level of understanding or learning gained from the text as a whole. Such judgements which relate to multiple pieces of information are referred to as ‘global judgements’ (Dunlosky & Lipko, 2007). Less frequently, judgement prompts are used which target smaller quantities of information, such as specific terms, and are referred to as ‘local judgements’ (Dunlosky et al., 2005; Nietfeld et al., 2005). It has previously been suggested that using prompts which target smaller quantities of information may result in perceptual judgements which are more closely related to performance on the experimental task (Dunlosky et al., 2005; Dunlosky & Lipko, 2007; Händel & Dresel, 2018; Händel et al., 2020; Koriat, 1995; Lefèvre and Lories, 2004). Should such an effect exist, prompting reader panellists to evaluate particular areas of concern within health texts would represent an easily implementable method to increase the informativeness of reader panellists’ judgements.

Contrary to expectations, previous research has yielded mixed findings concerning the influence of the scope of the judgement prompt on the relationship between perceptions of understanding and performance on a comprehension test. Studies have variously indicated that global judgements are more accurate than local judgements (Nietfield et al., 2005), local judgements are more accurate than global judgements (Connor et al., 1997), or that there is no difference between judgement types (Dunlosky et al., 2005; Schraw, 1994). Research has also suggested that any differences in accuracy associated with varying the scope of the judgement may be conditional on a retrieval attempt or a pre-judgement evaluation of perceived difficulty (Dunlosky, Rawson & McDonald, 2002; Dunlosky et al., 2005; Vossing and Stramov-Roßnagel, 2016). Given these inconsistent findings, the effect of judgement scope remains unclear. Moreover, the above research was conducted using methodology which differs from that most appropriate to investigate reader panellists’ judgements. The present research, therefore, will explore the potential utility of reducing the scope of the judgement prompts in the review of draft health documentation.

Secondly, the present research will investigate how information which varies in importance within the semantic structure of the text may impact the relationship between perceptual judgements and evidence of comprehension. It has previously been suggested that information within a text varies in its relative importance to the overall semantic structure of the text, with highly-connected ideas which feature at the greatest levels of representations capturing the main ideas, the overall gist and the topic of the text (Helder et al., 2019; Kintsch, 1994; Kintsch & van Dijk, 1978; Thorndyke, 1977; van den Broek et al., 2013; van den Broek & Helder, 2017; van Dijk & Kintsch, 1983). Support for this concept is provided by a body of research which suggests reading and recall processes are preferentially sensitive to semantic centrality (Helder et al., 2019; Kendeou et al., 2009; Nieuwland & Van Berkum, 2006; Stafura & Perfetti, 2014; van den Broek et al., 2013; van den Broek & Helder, 2017; Yeari et al., 2015; Yeari et al., 2017; Yeari & Lantin, 2020). In order to extract these main ideas, readers need to successfully engage in comprehension-building processes and form an inter-connected mental representation of the text (Kintsch & van Dijk, 1978; van den Broek et al., 2013; Yeari & Lantin, 2020). Given the potential importance placed on this goal, therefore, it may be the case that reader panellists' evaluations of health information may be influenced by their success in constructing and comprehending a set of coherent semantically central ideas following reading.

Importantly, if reader panellists' evaluations of written health information are indeed shaped by the desire to capture the main ideas of a text within a coherent relational hierarchy of meanings, the accuracy of reader panellists' judgements would be partly determined by the semantic centrality of the information assessed on the comprehension test. Specifically, subjective judgements would show a stronger relationship with performance on questions which target information central to the semantic structure of the text, whilst showing a weaker relationship with performance on questions which target information comparably peripheral to the semantic structure. For those interpreting reader panellists' responses, it would be valuable to know whether judgements preferentially inform of the capacity to comprehend the main points, rather than to understand more isolated pieces of information within a text, particularly where the latter may be critical for patient outcomes. Given the potential influence of semantic centrality on reader panellists' judgements, the present research will seek to clarify the relationship between subjective judgements and comprehension of information which varies in semantic centrality.

In sum, the present research aims to provide valuable insight for those who make use of reader panellist's judgements in the review of health information. This will be achieved through an item-level responses to comprehension questions obtained when the corresponding health-related texts were available during testing. Additionally presenting the traditional analytic approach within metacomprehension research, analyses of individual-level correlations, will permit an evaluation of the similarities and differences in the resulting findings. Critically, quantifying the strength and variability in the relationship between judgements of understanding of health-related texts and comprehension performance, alongside investigating the impact of judgement scope and semantic centrality, will furnish essential guidance regarding the effective use of reader panels.

7.2 Method

Secondary data were used in the analysis in the present chapter. The dataset consisted of a combined subset of the data obtained in Study 2 and Study 3 (Chapter 4 and 5, respectively), generating a dataset in which participants responded to the set same texts and comprehension questions. Any responses relating to comprehension questions which were revised in Study 3 were excluded (i.e., any questions which differed between Study 2 and 3 were removed to avoid potential conflation with any differences between studies arising from the experimental manipulation of interest). From the data collected in Study 2, observations included were judgement responses to the global judgement (text-level) prompt ‘Overall, how well do you understand the text?’ and responses to comprehension questions targeting semantically central and peripheral information within the text. From the data collected in study 3, observations included were judgement responses to local judgement (idea-level) prompts, corresponding comprehension questions targeting semantically central and peripheral information within the text, and responses to these comprehension questions. Only the data from participants who demonstrated sufficient engagement with the task were included in the combined dataset, determined as time spent on each stimulus text corresponding to a reading rate of up to 600wpm. Details of the participants, materials (including the construction of comprehension questions relating to semantically central and semantically peripheral information and global and local judgements) and procedure relating to the resulting combined dataset are provided below.

7.2.1 Participants

The secondary data used in the present study consisted of responses from 396 participants, with 215 female, 180 male and 1 non-binary participants, recruited via Prolific (Prolific.ac). Participants’ age ranged from 18 to 77, with an average of 40.54 (*SD* = 14.16). As judgement condition was a between-subjects variable, participants were classified as either in the global or local judgement condition. A total of 225 participants provided responses in the global judgement condition, and 171 participants provided responses in the local judgement condition. As semantic centrality of the comprehension question was a within-subjects variable, participants in both judgement conditions provided responses to the central and peripheral comprehension questions. Participants were compensated for completing the study: participants in the local judgement condition received £5, while participants in the global judgement condition received £6 (due to participants in the latter group also completing individual difference measures which are not analysed here).

7.2.2 Materials

Health Texts.

Sampling Procedure. Text sampling was done in a two-step approach: 10 texts were initially selected and text features were evaluated, followed by the selection of a further three. An opportunity sample of ten sources of online health information was collected via searching the web. To avoid inaccurate health information, only sources which declared the involvement of medical expertise in the production of the information were included. Sources which were fully pay-walled were excluded. From each source, one text topic was selected.

To select text topics, 10 conditions were randomly identified, using a random letter and number generator, from the A-Z of rare diseases listed on rarediseases.org. Topics were selected by iteratively randomly selecting a topic from the list and then verifying its suitability. Suitability was assessed according to whether the topic was appropriate (not likely to be an embarrassing or distressing condition), featured on the target source website, and whether there was sufficient text to create a 200-300 word stimulus text.

After the selection of 10 text topics, stimuli were constructed from the online information (full description in the following subsection). Descriptive information for the constructed stimuli was then generated using the Coh-Metrix 3.0 online tool (Graesser, et al., 2004) to evaluate text features of interest. This roughly indicated there was a cluster of seven texts which scored more highly on readability measures, compared to three lower readability texts. To provide more opportunities to observe lower judgements of understanding, an additional text was obtained from each of the three sources from which the lower readability texts were constructed, selected using the sample strategy described above, producing the full set of 13 health-related texts.

Stimulus Construction. For each health topic, sections of text from the source webpage were taken to produce 200-300 word texts. Links, images and formatting were removed, but heading and subheadings were preserved. Sections which were predominantly lists, or discussed aspects likely to cause distress, were omitted. The coherence of the resulting text was checked, by rereading, to ensure that flow and meaning was preserved. American English spellings were changed to British English. Descriptive information for the texts is presented in Table 7.1 and the full texts are provided in Appendix I.

Table 7.1

Descriptive Information for the Health Text Stimuli

Topic	Words	Flesch Score	Fleisch-Kincaid
Porphyria	273	61.3	8.47
Scleroderma	257	55.8	8.51
Neutropenia	214	9.4	15.27
Seborrhoeic dermatitis	228	59.6	8.49
Erysipelas and Cellulitis	241	51.4	9.24
Cholestasis of pregnancy	268	60.5	7.48
Dengue Fever	286	51.1	9.36
Bilateral renal agenesis	243	60.7	8.35
Zollinger-Ellison syndrome	247	31.9	12.65
Myasthenia gravis	252	29.8	14.32
Coxiella burnetii infection	231	8.03	16.24
Sialadenitis	257	42.79	10.07
Brucellosis	285	38.38	12.96

Note. Flesch score refers to the Flesch Reading Ease score. Flesch-Kincaid refers to the Flesch-Kincaid Grade level. Calculated using the online tool Coh-Metrix 3.0 (Graesser, et al., 2004).

Comprehension Questions. For each text, six comprehension questions were constructed: three targeting semantically central information and three targeting semantically peripheral information. To construct these questions, information which varied in semantic centrality was identified from the texts. Given the variable approaches in defining and quantifying centrality reported in previous literature, an attempt was made to create a procedure to identify central and peripheral elements with greater objectivity and control. The approach incorporated ideas from multiple sources, concerning the nature and features of central and peripheral information (e.g., Helder et al., 2019; Kintsch, 1988; Thorndyke, 1977; van den Broek et al., 2013; van den Broek & Helder, 2017; van Dijk & Kintsch, 1983). In the development and application of the approach, however, it was recognised that i) the process was not an exhaustive measure of the ways in which an idea in a text may be more or less central, and ii) a degree of subjectivity was unavoidable (owing to the difficulty in operationalising some concepts, such as connections and elaborations).

For each text, information was divided into coarse idea units, expressing a single idea, which were then used to construct a semantic map of the connections between ideas in the text. The semantic maps were jointly constructed by two researchers, with disagreements resolved through discussion. From the semantic map, information which was more versus less-well connected, representing semantically central versus peripheral information, was selected (see Appendix J for details). As not all information was amenable to question construction, some idea units were not considered suitable for use. For example, some idea units contained only numerical information (e.g., “40,000 are affected”) which presented a challenge in avoiding testing only surface level processing. Following this procedure, three semantically central and three semantically peripheral idea units were identified for each text. The full procedure is detailed in Appendix J.

After identifying 78 idea units, a small-scale study was conducted to collect ratings of centrality. This information was obtained to examine the relationship between the selected idea units and readers' perceptions of centrality. Full details of the method of obtaining the centrality ratings data are provided in Appendix Q section Q.1. The relationship between the identified idea units and readers' perceptions of centrality was evaluated through data visualisation. Overall, idea units identified as central were typically rated as more semantically central, on average, than those identified as peripheral. However, substantial within-classification variability was observed, with the ratings of some idea units incongruous with their identification (see Figure Q.1 in Appendix Q).

Multiple-choice questions were developed to assess understanding of information of the 78 idea units identified. It was considered that assessing understanding of the idea unit itself, rather than of an inference generated on the basis of it, was of primary relevance. Idea units captured individual pieces of information which, while they could be used to form inferences, were not derived from inferences themselves. Questions were designed to test comprehension of the information corresponding to the idea unit which was explicit in the text, while avoiding surface overlap. Identifying the correct answer required matching a semantically accurate, but lexically dissimilar, response option with the information provided in the text. Three distractors were created for each question item. Distractors primarily consisted of response options with high lexical overlap with the text, likely misconceptions, or semantic near-misses. Constructed questions and response options were separately reviewed by three experimenters and altered until all experimenters were satisfied. Alterations were made to 11 comprehension questions based on the data collected in the global judgement condition. To avoid introducing ambiguity into the dataset, questions which varied between judgement conditions were removed from the dataset prior to analysis. The final set of 67 comprehension questions with response options, administered in both the global and local judgement conditions, is provided in Appendix Q section Q.2.

Judgements of Perceived Understanding. In the global judgement condition, the prompt 'Overall, how well do you understand the text?' was used to elicit a judgement of perceived understanding for each text. In the local judgement condition, six prompts were designed per text to elicit judgements of perceived understanding, corresponding to each of the questions on the comprehension task for a given text. Each local judgement prompt was prefaced with the wording 'How well do you understand ...', followed by the target information. Prompts were constructed with consideration of the potential influence they may have on participants' response processes. To reduce the potential for a correct response on the comprehension task to be generated by recall of the prompt or reduced plausibility of response options, prompts were worded to avoid stating the correct answer to the subsequent comprehension question or invalidate distractors. Simultaneously, prompts were worded to limit ambiguity in the information offered for evaluation and to ensure that this information corresponded to the same information tested in the subsequent comprehension questions. The judgement prompts were separately reviewed and altered until three experimenters were satisfied with the wording. Reflecting the exclusion of responses to comprehension questions which were altered during the data collection between judgement conditions, responses to 11 prompts were removed from the dataset prior to analysis. The resulting set of 67 local judgement prompts is provided in Appendix Q section Q.3.

For both global and local judgements, ratings were captured using 7-point rating scale. The rating scale was presented as an unmarked line with a moveable slider to be dragged left or right to indicate judgement (the slider 'snapped' to the value closest to one of the seven ratings). The labels 'not at all well' and 'extremely well' were presented on the left and right of the scale, respectively.

7.2.3 Procedure

For individuals recruited to both the global and local judgement conditions, participation in the study commenced when participants responded to the online invitation on Prolific.co. All tasks were computer-based and presented via Qualtrics. Participants provided consent and reported their age and gender prior to completing the tasks. Participants first read each health text in the experiment, presented in one of four randomised orders, and made judgements of comprehension immediately following each text. Participants in the global judgement condition were prompted to provide a single global rating of perceived understanding, shown without the text, after reading. Participants in the local judgement condition were presented with the six local judgement prompts, shown together on screen without the text, after reading. Participants were invited to take a break if needed after reading and judging all the texts. Participants were then presented with the comprehension questions for each text, with the text present, in the same order that the texts were presented for reading and judging. Participants in the global judgement condition were invited to complete a second session, approximately one week later, which included a battery of individual difference measures not analysed in the present study. After completing the tasks, all participants were fully debriefed.

7.3 Results

To furnish clear guidance on the effective use of reader panels, the strength of the relationship between individuals' judgements of perceived understanding of health information was analysed, alongside the potential influence of the scope of the judgement prompt and the semantic centrality of the information within the text. As discussed, the applied context of reader panel judgements motivates an item-level analyses of responses. However, given the novelty of this approach, the traditional correlational approach was also conducted to permit an evaluation of findings under both analyses. In the remainder of this section, descriptive statistics are first discussed, followed by the presentation of the findings relating to the correlational and logistic regression analyses.

Descriptive data for the judgment and performance measures, by experimental condition, are provided in Table 7.2. The magnitude of judgements in the global condition was slightly greater, on average, than in the local judgement condition. Performance on the comprehension task was similar across judgement conditions and question types. However, performance was lower, on average, for participants in the global judgement condition responding to semantically peripheral comprehension questions.

Table 7.2

Average Judgement Magnitude and Proportion Correct Test Performance, with Standard Deviations in Parentheses, by Experimental Condition

		Performance	
Condition	Judgement	Semantically Central	Semantically Peripheral
Global Judgement	5.43 (1.57)	0.69 (0.15)	0.58 (0.14)
Local Judgement	4.74 (1.73)	0.68 (0.14)	0.72 (0.19)

7.3.1 Gamma Correlation Analysis

For each participant, a set of judgement-performance dyads were used to calculate individual-level gamma correlation coefficients (note that repeating this analysis using Pearson correlations yielded the same results). For participants in the global judgement condition, correlations were calculated using 13 judgement-performance dyads, each consisting of a text-level judgement and the corresponding sum total of questions correctly answered on the comprehension task for each of the 13 texts. The coefficient could not be calculated for one participant, due to a lack of variation in their judgements, and was excluded from further analysis. For participants in the local judgement condition, correlations were calculated using 78 judgement-performance dyads, each consisting of an idea-level judgement and the binary observation of whether or not the corresponding comprehension question was answered correctly, with six of dyads associated with each of the 13 texts.

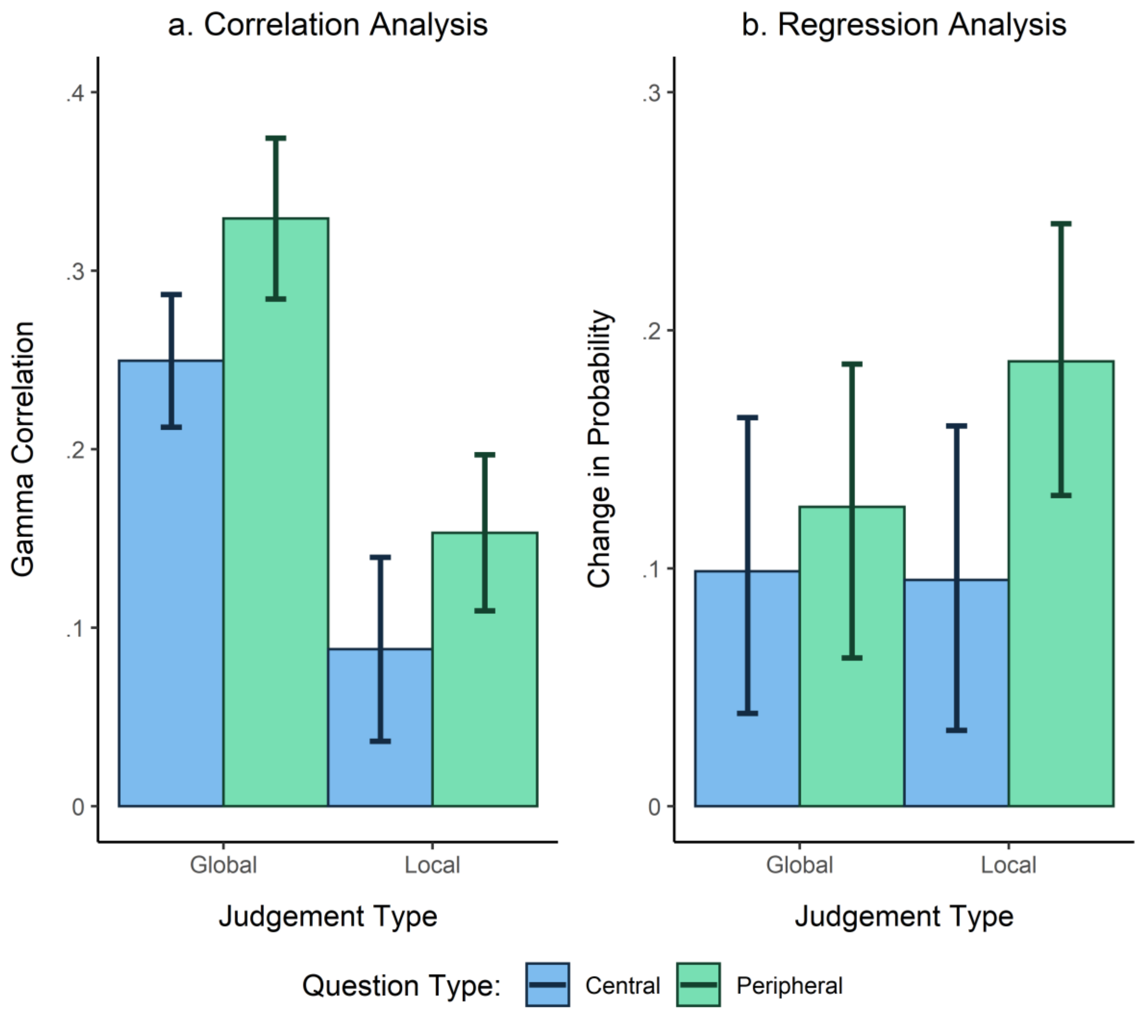
Participant-level gamma correlations ranged from -1 to +1 across experimental conditions, indicating considerable variability across individuals in the strength of the association between judgement and performance measures (see Figure Q.4.1 in Appendix Q for distributions of participant-level gamma correlations by experimental condition). The individual estimates for participant-level gamma correlations were typically associated with considerable uncertainty, with the standard error ranging between 0.23 and 0.32 across experimental conditions on average.

A mixed 2x2 ANOVA was conducted to statistically examine differences in the average participant-level gamma correlation by experimental condition, with judgement type as a between subject factor and question type as a within subject factor. The main effect of judgement type was significant: $F(1, 393) = 48.45$, $p < .001$. The main effect of question type was also significant: $F(1, 393) = 11.70$, $p < 0.001$. The interaction between judgement type and question type was not significant ($p = 0.73$).

The ANOVA indicated that the average gamma correlation was greater given global judgements ($M = .29$) compared to local judgements ($M = .12$), regardless of question type. In addition, regardless of judgement type, the average gamma correlation was greater for semantically peripheral questions ($M = .25$) compared to semantically central questions ($M = .18$). The average participant-level gamma correlation for each experimental condition is shown in Figure 7.1a, with error bars showing the 95% confidence interval for the mean, grouped by judgement type (global on left; local on right) and coloured by question type (central shown in blue; peripheral shown in green).

Figure 7.1

Estimated Differences Between Experimental Conditions According to the Correlational Analysis (Left) and Logistic Regression Analysis (Right)



Note. The left panel plot (a.) shows magnitude of average participant-level gamma correlation by experimental condition. Error bars shown in the left panel represent the 95% confidence interval for the mean. The right panel plot (b.) shows the average estimated increase in the probability of observing a correct response given a maximum judgement of understanding, compared to a minimum judgement of understanding. Error bars shown in the right panel represent the 95% credible interval for the mean. Estimates are grouped by judgement type (global on left; local on right) and coloured by question type (central shown in blue; peripheral shown in green).

7.3.2 Multilevel Logistic Regression Analysis

The combined observation-level data from participants in both judgement conditions was used to fit the multilevel regression model, with each row of the data frame consisting of a participant's judgement, performance, and the respective condition identifier. Observations of judgement of perceived understanding, judgement condition and question type were entered into the model, in addition to the interactions between these variables. Multilevel variances in the intercept were estimated for participants, texts and questions. Participant-level variances in the effects of judgements and question type were also included. No participant-level variance in the effect of judgement type was estimated as this was not a repeated measures variable. Full details of the model fitting, in addition to details of the resulting full model fit, are provided in section Q.5 of Appendix Q.

The results of the logistic regression are illustrated in Figure 7.1b, which shows the expected change in the probability of observing evidence of understanding in each condition when participants report high, compared to low, levels of understanding, with error bars displaying the 95% credible interval. The regression model indicated that judgements were reliably predictive of comprehension outcomes in all conditions. This can be seen in Figure 7.1b as the credible interval for the average fitted effect in each condition does not overlap with zero, indicating a positive effect of higher judgements of understanding on the chance of demonstrating comprehension. Overall, however, no effects of judgement scope, question type or the interaction between these variables were found to be robust. This can be observed in Figure 7.1b as the credible intervals for the average effect in each condition are overlapping.

Specifically, in the global judgement condition, the estimated effect of judgements (magnitude of the beta slope) was highly similar for both semantically central comprehension questions ($\beta = 0.10$) and semantically peripheral comprehension questions ($\beta = 0.13$), indicating no reliable effect of semantic centrality. This comparability may be sensitive to the inclusion of question-level variance parameters using item-level scoring, however (see section Q.6 in Appendix Q for the full details of a supplementary regression analysis omitting question-level variance parameters using sum scoring, i.e., a binomial model). In the local judgement condition, a larger difference in the effect of judgements of understanding was estimated for semantically central comprehension questions (β = 0.10) compared to semantically peripheral comprehension questions (β = 0.19). Although this difference was reliable according to a 90% credible interval, the difference did not reach the 95% threshold typically considered as robust evidence for an effect. Conclusive support for an interaction between the scope of the judgement prompt and the semantic centrality of the information was therefore not found.

With respect to individual-level variability in the estimated relationship between judgements of understanding and the likelihood of demonstrating comprehension, the regression model indicated that participants were highly similar. Across all conditions, participants showed a positive association between perceived and assessed understanding (see Figure Q.4.2 in Appendix Q for distributions of participant-level increases in probability given the maximum judgement of understanding). No participants were found to show a negative association between judgements and performance on the task.

7.4 Discussion

The present research aimed to address the question of how reader panellists should be instructed, and their judgements interpreted in, order for their input to be effective in helping to produce understandable patient information. Analyses of the relationship between judgements of perceived understanding and comprehension performance on health-related texts were conducted to examine whether reader panellists judgements can inform of their capacity to understand the information presented to them. In addition, to equip those who make use of reader panels with greater insight, the present research investigated the potential effects of judgement scope and semantic centrality. Two statistical approaches were used to analyse the data, permitting a comparison between the traditional approach used within metacomprehension research (correlational analysis) and an alternative approach considered appropriate for the present research (multilevel logistic regression analysis).

Overall, in both the correlational and multilevel logistic regression analyses, the relationship between perceived judgements of understanding and performance on the comprehension task was found to be weakly positive. This effect was replicated across all experimental conditions, with higher judgements of understanding showing a limited association with greater evidence of comprehension. These findings are consistent with those of metacomprehension studies which indicate that judgements of learning and understanding are poorly predictive of test performance (Dunlosky & Lipko, 2007; Lin & Zabrucky, 1998; Maki, 1998; Prinz et al., 2020a; Yang et al., 2022). In addition, a similar pattern of differences for the influence of semantic centrality was observed in both the correlational and regression analyses: showing a tendency for judgements to be more predictive of performance on questions targeting semantically peripheral information compared to semantically central information.

However, the findings yielded under the two analytic approaches diverged in a number of ways. While both approaches indicated differences consistent with a main effect of semantic centrality, this effect was not reliable in the regression analysis. The main effect of judgement scope found in the correlational analysis also failed to replicate in the regression analysis. Conversely, evidence for a potential interaction between the scope of the judgement and semantic centrality found in the regression analysis was not replicated in the correlational analysis. In addition, the two analyses produced markedly different findings with respect to individual variability. The correlational analysis indicated that participants may variously range from showing a perfectly negative or perfectly positive association between their judgements and comprehension performance. In contrast, the regression analysis suggested that there are only small differences between participants in the weakly positive relationship, with no negative associations estimated for any participant.

In considering the divergence in findings between the two analytic approaches, the two analytic approaches applied in the present research differ critically in how uncertainty is incorporated at the question and individual level. Under the traditional approach, in calculating non-parametric participant-level correlations (i.e., gamma correlations), ordinal differences in performance across a summed set of questions are assessed against judgement magnitudes. For parametric correlations (e.g., Pearson's correlations; and similarly in binomial regression), in producing the sum scores which capture comprehension performance, individual questions are treated as independent and identically distributed, with a fixed and equal level of difficulty. In submitting the participant-level correlations into an ANOVA to evaluate differences in the average correlation between experimental conditions, each participant's correlation (which is a sample-based estimate with associated uncertainty) is treated as a fixed value with zero variance. In contrast, under the multilevel logistic regression approach, in calculating the relationship between judgements and performance, responses are constructed as being derived from a population of questions with a fixed average difficulty, but with idiosyncrasies between questions (Kulesz et al., 2016). Further, in calculating differences in the average relationship between judgements and performance between experimental conditions, within the multilevel logistic regression model, the participant-level relationship between judgements and performance is itself treated as an estimate with associated uncertainty.

While differences in results may be attributed to the steps involved in alternative approaches outlined above, the characteristics of the participants, experimental materials, and the methodological approach can produce tangible differences in the extent to which the properties of a resulting dataset accord with the statistical assumptions underpinning any particular analysis (Rohrer & Arslan, 2021). The divergence in findings observed between the analytic approaches applied here, therefore, may not be reproduced in other contexts. For example, greater concordance between analytic approaches may be found between studies in which the text is absent during the comprehension test or where there are fewer stimulus texts. This view accords with research which suggests that the relationship between metacomprehension judgements and performance may be influenced by aspects of the methodology (Ozuru et al., 2013; Prinz et al., 2020a; Paulewicz et al., 2020; Vuorre & Metcalfe, 2022). Differences between the findings of the two analytic approaches observed here, therefore, cannot be used justify the suitability of any specific analysis of any particular dataset. However, given the context in which reader panellists' judgements are made and used, the multilevel logistic regression is considered most appropriate analysis to interpret presently. This item-level analysis most closely corresponds to the conceptualisation of successful comprehension of health-related information, in which patients are able to locate and understand elements of the text, allowing them to make informed decisions about their health (European Commission, 2009; Institute of Medicine, 2004). The remainder of this section, therefore, discusses the findings of the regression analysis alongside the implication for those who make use of reader panels.

Overall, the results of the regression model indicate that reader panellists' judgements of understanding can be considered a valid tool in assessing the likely comprehensibility of draft health documentation. However, the strength of the association between judgements and performance is weak: across experimental conditions, individuals were only approximately 1/8th more likely to successfully use written health information to demonstrate understanding on texts rated at the highest level of comprehensibility compared to the lowest level. While high judgements are associated with greater capacity for showing understanding, therefore, elements of the information may remain poorly understood. Nevertheless, lower judgements are more likely to be observed on texts which are associated with lower comprehension. Such judgments, therefore, provide a useful indicator that the text may require revision.

No robust differences between the judgement scope and semantic centrality conditions were found in the regression analysis, with the 95% credible intervals for each effect estimated to be overlapping. The estimated relationship between judgements and performance were highly similar for global judgements of semantically central and peripheral information, and for local judgements of semantically central information. This estimated equivalence demonstrates that eliciting judgements of understanding for the entire text and for specific core ideas produces judgements which are similarly predictive. Instructing reader panellists to specifically evaluate key ideas will, therefore, generate judgements which are comparable in interpretation to judgements of understanding made of the whole text. Global judgements of understanding were also similarly predictive of the comprehension of semantically peripheral information, indicating no discrimination in global judgements for information varying semantic centrality. Instructing reader panellists to evaluate draft patient information as a whole, therefore, provides insight into the likely comprehensibility of both key ideas and isolated pieces of information. Given this pattern of results, it may be speculated that the process of generating global judgements and local judgements of semantically central information is closely related, though the association with comprehension performance remains weakly positive across these circumstances.

In contrast, while the estimated difference was not robust, the analysis indicated that idea-level judgements may be particularly predictive of demonstrating comprehension on isolated pieces of information within the text. This provides some evidence that targeted, idea-level judgements may be more predictive of demonstrating comprehension of specific pieces of information which require less integration across the text. However, since the credible interval for this effect overlapped with those of the other conditions, further investigation would provide greater confidence in the magnitude of this effect. Moreover, given that the ANOVA yielded a different pattern of results (main effects of both judgement scope and semantical centrality), additional research is required to more fully examine how the veracity of the findings may be impacted by alternative analytic approaches.

The limited indication that local judgements are more predictive of understanding semantically peripheral information suggests that the judgement process may differ under these circumstances. Individuals may be more accurate when specifically asked to evaluate their understanding of text which requires lower engagement in integrative comprehension-building processes when the document is available for reference during the comprehension assessment. Such an effect would substantiate existing guidance that advises reviewers are asked about discrete, self-contained aspects within patient information leaflets to assess understanding (European Commission, 2009). For those making use of reader panels, therefore, in reviewing information which requires limited higher-level processing to achieve adequate comprehension, using targeted judgement prompts may well provide greater insight into the likelihood that such information is understood.

Importantly, the regression analysis indicated that the weakly positive relationship between judgements of understanding and comprehension performance was consistent across individuals This limited variation in the strength of the association between individuals was observed in each experimental condition. Although the correlational analysis conversely indicated high variation between individuals, each individual-level correlation estimate was itself associated with high-levels of uncertainty. Considering this uncertainty, therefore, the underlying individual-level correlations may be more similar than the point-estimates alone would indicate. While additional research would be beneficial in clarifying the source of the variability in individual-level correlations, the findings of the regression analysis suggest that judgements of perceived understanding similarly inform of the likely comprehensibility of a health document across individuals. This insight is highly valuable for those interpreting reader panellists' judgements, as the quality of evaluations do not appear to substantially differ between panellists. For those recruiting reader panellists, this finding indicates that introducing selection-criteria may not improve the quality of feedback provided by panellists.

Although the present research is primarily concerned with providing insight to those who make use of reader panels, the findings may also be useful in helping to further understanding of the cognitive processes underlying judgements of comprehension. Existing theoretical accounts of metacomprehension judgements have been developed upon research with methodology which differs from the present research. Given such differences, existing theories are arguably unable to appropriately accommodate the present findings. However, it is reasonable to speculate on how existing theories may be extended to the judgement processes employed by reader panellists in the evaluation of health information. Inferential accounts of metacomprehension judgements have previously suggested that individuals rely on cues as the basis of their judgements of understanding (Dunlosky, Rawson & Hacker, 2002; Griffin et al., 2009; Koriat, 1997; Zhao & Linderholm, 2008; Wiley et al., 2016). Such accounts offer explanatory utility for the weak association observed between judgement and performance in the present research: perceptual judgements are imperfect predictors of text comprehension as individuals are not afforded direct access to the products of comprehension to evaluate the quality of their understanding.

Accounts of metacomprehension judgements which have emerged from research conducted within an educational context typically suggest heterogeneity in the cues selected when students evaluate their learning and understanding of instructional materials (Griffin et al., 2013; Griffin, Mielicki & Wiley, 2019; Koriat, 1997; Zhao & Linderholm, 2008).

However, the notable similarity in the relationship between judgements and performance observed in the present research indicates a fundamental similarity in the judgement processes across individuals in the context of evaluative judgements of health information. The availability of the text during the assessment of comprehension may be fundamental to the observed similarity between individuals. In contrast to metacomprehension research which does not make the text available during testing, variability in the efficiency of search behaviours and different levels of processing may feature more strongly in the measure of understanding obtained in the present research (Ardoin et al., 2019; Cataldo & Oakhill 2000; Ozuru et al., 2007; Schroeder, 2011). To better understand the cognitive processes which underpin and link reader panellists' judgements and the capacity to use and comprehend written health documents, identifying what drives the similarity in the predictive utility of reader panellists' judgements would be a productive goal for future research.

Importantly, further research is also required to address two limitations of the present research. Firstly, no a priori analysis of the statistical power or sensitivity of the present research was conducted prior to conducting the analyses. As a result, the present research may be underpowered, meaning that the effects are more likely to be estimated with lower accuracy and precision (Johnson et al., 2015; Maxwell et al., 2008). To ensure robust estimates of effects are obtained and differences between experimental conditions can be effectively assessed, further research with adequate power is required. Secondly, in the present research, comprehension performance was generally high: participants responded to 65% of the comprehension questions correctly, on average, with no participants scoring below 25%. It is not clear, therefore, whether the findings are representative of lower ability readers. Researchers have previously suggested that, amongst lower ability readers, metacomprehension judgements may be less predictive of understanding (Griffin et al., 2008; Ozuru et al., 2012). Although these individuals may be less likely to volunteer as reader panellists, clarifying whether such an effect exists is important. Low levels of literacy are associated with poorer health (Weiss et al., 1992) and attempts to identify a lack of understanding in patients with low literacy could be further frustrated by assuming that the results observed here (indicating a positive relationship between judgements and performance) similarly apply to such individuals.

In conclusion, the present research suggests that reader panellists' judgements provide valid but limited insight into the likely comprehension of draft documentation for patients. Overall, the relationship between judgements of perceived understanding of health information is not strongly predictive of comprehension outcomes. Whilst low judgements of understanding are likely to indicate problems experienced when attempting to comprehend the text, individuals may still have incomplete understanding despite reporting high levels of comprehension. Limited variation between individuals indicates that the informativeness of judgements is unlikely to substantially vary across reader panellists. However, this relationship may not generalise to lower ability readers, who were not well-represented in the participant sample. Targeted judgement prompts may provide greater insight into the likely comprehension of information which requires limited high-level comprehension-building processes, although additional research is necessary to verify this. Overall, whilst reader panellists' judgements of understanding can act as a proxy for comprehension, practitioners who make use of these judgements should consider the whether the insight yielded is sufficient for their purposes.

8. Appendices

Appendix A: Stimulus texts used in Pilot 1

1) Femoral Hernia

Femoral hernia

A femoral hernia is a loop of intestine, or another part of the abdominal contents, that has been forced out of the abdomen through a channel called the 'femoral canal' - a tube-shaped passage at the top of the front of the thigh. The loop is usually only the size of a grape.

A femoral hernia can cause serious medical problems if left untreated, even if there are no troublesome symptoms to begin with. Treatment is by an operation to return the herniated intestine to its proper place and close the weakness in the abdominal wall.

About femoral hernias

The femoral canal, through which a femoral hernia is squeezed, is next to the point where the blood vessels and nerves pass from the abdomen into the leg. It is a potential weak spot in the abdominal wall.

Intestine (bowel), or the tissue that covers it, is more likely to be forced out through the femoral canal if a weakness already exists. Increasing the pressure inside the abdomen, by activities such as standing up, coughing or straining can then trigger a hernia. Other factors that make a femoral hernia more likely to develop include:

- Being very overweight (obese)
- Having a smoker's cough
- Constipation
- Carrying or pushing heavy loads

Femoral hernias tend to occur in older people. It also appears that pregnancy may weaken the abdominal tissues, making femoral hernias more common in women who have had one or more pregnancies.

Symptoms

A femoral hernia causes a grape-sized lump in the groin, although this is not always easily noticeable.

If the hernia can be manually pushed back into the abdomen it is referred to as 'reducible'. However, usually this is not possible and the hernia is effectively stuck in the canal. This is an 'irreducible' hernia and is a potentially dangerous condition. The blood supply to the herniated tissue can become crushed within the canal, cutting off its source of oxygen and nutrients. This is known as a strangulated hernia and emergency surgery must be performed to release the trapped tissue and restore its blood supply. A strangulated hernia is very painful and tender to the touch.

Once a hernia has formed it is important to seek a doctor's advice. A truss (a type of corset designed to hold in a hernia) should not be used for a femoral hernia as it can encourage the hernia to become strangulated.

Treatment

All femoral hernias need to be treated surgically as they have a high risk of becoming strangulated.

A femoral hernia repair is routinely performed as a day case, without the need for an overnight stay in hospital. The type of anaesthesia will depend on the exact operation and the preferences of the surgeon and patient. Femoral hernia repairs are routinely carried out under general or regional anaesthesia (where just the area being treated is anaesthetised).

The operation

The surgery is generally performed through an incision about 10cm long either over the hernia itself or on the lower abdomen. The procedure involves opening up the femoral canal, returning the loop of intestine or intestinal covering back to the abdomen, and then patching up the canal to repair the defect that let the hernia through in the first place. The top of the femoral canal may be reinforced by a mesh made of a synthetic material that does not irritate the body.

Laparoscopic surgery, also known as 'keyhole' or 'minimally invasive' surgery, may be used. If the hernia has become strangulated, and part of the intestine damaged, the affected segment of intestine may need to be removed and the two ends of health intestine connected. This is a more complex surgery and requires a longer stay in hospital.

What are expected results after having Laparoscopic Hernia Repair surgery versus having an open abdominal surgery?

- Decreased postoperative pain
- Shortened hospital stay
- More rapid return to bowel function
- More rapid return to work
- Minimally sized incisions with a better cosmetic result

What are the risks of having Laparoscopic Hernia Repair surgery?

As with any surgery there are risks. The risks of one of these complications is no greater than if the surgery were done with the open technique. Complications that can occur are:

- Bleeding
- Infection involving the wound, blood or abdomen
- Injury to surrounding organs such as the bladder, intestines, blood vessels, nerves or the spermatic tube that goes to the testicles (males)
- Difficulty urinating following surgery may occur and a temporary catheter may be ordered to drain the bladder
- Numbness and pain in the groin region may require an open surgery technique
- Even though a hernia may be repaired, it may return

You should ask your surgeon any questions you have in regards to the risk and benefits of the procedure.

What happens if Surgery cannot be performed by Laparoscopic Technique?

Sometimes it is not possible for the surgeon to use the laparoscopic technique because it may be difficult to see or handle tissue safely. The surgeon decides to perform an open procedure either before or during the surgery. The surgeon may decide to convert the laparoscopic surgery to an open procedure in certain situations and for patient safety. Though very infrequent, when conversion to an open technique occurs, it should not be considered a failure of the procedure. Factors that might increase the possibility of changing to an "open" procedure are obesity, previous abdominal surgery causing dense scar tissue, inability to see organs or bleeding during surgery.

After the procedure

If the operation is a day case, most people go home once they have recovered from the anaesthetic. Anyone who has a general anaesthetic will need to arrange for a friend or relative to drive them home and stay with them for the next 24-hours.

A general anaesthetic can temporarily affect co-ordination and reasoning skills, so people are advised to avoid driving, drinking alcohol or signing legal documents for 24-hours afterwards.

Before discharge, a nurse will advise you about caring for stitches and bathing, and arrange a date for a follow-up appointment (about six weeks later).

Once home, painkillers should be taken as advised by the doctor or nurses. Whether recovering from open or keyhole surgery, it will be necessary to take it easy for the first two or three days. The surgeon will give specific advice about resuming normal activities. In general people will be able to move around freely but should avoid strenuous exercise and lifting for at least the first few weeks.

Most people continue to experience some discomfort for a few weeks after the operation, but this will gradually settle.

Deciding to have hernia repair

A femoral hernia needs to be treated to prevent strangulation, and it will not get better by itself. Surgery is the only cure for a hernia.

A femoral hernia repair is generally a safe surgical procedure. However, in order to give informed consent, anyone deciding to have this operation needs to be aware of the possible side-effects and the risk of complications.

Side-effects

Side-effects are the unwanted but usually temporary effects of a successful procedure. Examples include feeling sick as a result of the general anaesthetic or painkillers.

Complications

Complications are unexpected problems that can occur during or after the operation. Most people are not affected. The main possible complications are an unexpected reaction to the anaesthetic, excessive bleeding, infection or developing a blood clot, usually in a vein in the leg (deep vein thrombosis). To help prevent this, most people are given compression stockings to wear during the operation.

Specific complications of a femoral hernia repair are uncommon but can include accidental damage to internal organs, which could require a larger incision to repair. There is also a risk of abdominal bruising, although this usually settles without treatment.

2) Basal cell carcinoma text

What are basal cell carcinomas (BCCs)?

Basal cell carcinomas, sometimes called 'BCC' or 'rodent ulcer', are a type of skin cancer. More than 80,000 cases are diagnosed per year, making them the most common cancer in the

UK. BCCs tend to affect the over 50s. Occasionally they are found in patients in their 20s, 30s and 40s.

Why do BCCs occur?

BCCs arise due to too much sun exposure in people with fair skin. Even casual sun exposure from day-to-day activities is enough for some people to develop BCCs. They are more likely to occur on body areas that catch the sun, such as the face, scalp, neck, back and chest. It is important to understand that there is a lag period of several years (sometimes decades) between sun exposure and developing BCCs.

What is the outlook?

Very good, as treatment is effective and usually provides complete cure. BCCs are different from many other cancers as it is extremely rare for them to spread elsewhere in the body. They do not usually cause ill-health or shorten life but will continue to grow unless they are treated. Lesions that have been neglected for many years can be harder to cure as they grow under the skin into nearby structures such as nerves, muscle and bone.

What do BCCs look like?

Some BCCs appear as flat scaly red patches while others are pink spots or lumps. There may be a small sore or ulcer which scabs or bleeds and will not heal. Some BCCs can be very subtle and resemble a scar or a dent. Most lesions are painless, and are often only noticed if they scab or bleed.

How do BCCs grow?

BCCs grow at the site at which they have arisen. They usually grow slowly over several months or years. They do not go away. BCCs have roots around and below the visible lesion. The roots can only be seen with a microscope. The lesion enlarges as the roots expand, similar to a weed. If the roots are not treated, then the BCC will come back - just like a weed. This is an important concept to understand.

How are BCCs diagnosed?

BCCs can be diagnosed from their appearance by a trained professional. Sometimes a skin biopsy is required to confirm the diagnosis - this is when a small sample of a lesion is removed for testing.

Do BCCs need to be treated?

Yes. BCC is a cancer and so treatment is nearly always essential. If not treated, BCCs will continue to grow and damage the skin and possibly nearby structures.

How are BCCs treated?

The best treatment for you depends on your age, and health, and the site, size and number of BCCs you have. Treatments are designed to treat the visible growth and surrounding roots as well. Your Dermatologist will discuss the possible treatment options with you, including:

Surgery

This is the most common way of treating BCC. This involves cutting away the lesion together with some surrounding skin. A minimum safety margin of 4 to 6 millimetres of skin around a lesion is removed to make sure all the roots are also removed. The area is usually stitched together though sometimes a skin graft is needed. Most surgery will be carried out using a local anaesthetic (this means you are awake and injections are used to numb the area), and as a day-case procedure. Sometimes a general anaesthetic will be needed and so a short stay in hospital will be necessary.

Radiotherapy

Radiotherapy treats cancer by using high-energy X-rays which destroy the cancer cells while doing as little harm as possible to normal cells. For some people this may be a more appropriate treatment than surgery. In this case you will be referred to a clinical oncologist (a consultant specialising in using radiation to treat cancer) who will discuss the treatment with you in detail. Radiotherapy is also sometimes used after surgical excision to help ensure that the cancer does not return.

Some superficial BCCs may also be treated by:

- Curettage and cautery - the lesion is scraped away
- Cryotherapy - the lesion is frozen using liquid nitrogen
- Cream - an anti-cancer cream is applied regularly at home

What happens after treatment?

Most patients will not need to be followed up in the clinic after treatment. However you should check the treated area each month, as there is a very small chance that the BCC may return. It would look similar in appearance to the original BCC. It is estimated that 1 in 20 BCCs may return in the 5 year period following treatment. It is possible that you may develop a new BCC somewhere else. You should therefore check your skin regularly - particularly on the scalp, face and neck. You should see your GP if you are worried about a new lump or skin lesion, if it has been present for more than 6 weeks, is getting bigger, scabs or bleeds.

How can I prevent further BCCs?

You can also take some simple precautions to help prevent further skin cancers developing:

- Do not allow yourself to sunburn
- Do not try to get a suntan - going out in the sun with specific intention of going brown will increase your risk of skin cancer
- Cover up on a bright day. Protect the skin with clothing, including a hat, T-shirt and UV protective sunglasses
- Avoid strong sunlight. Spend time in the shade when it's sunny particularly between 11:00 and 15:00
- Use a 'high protection' sunscreen of at least SPF 30 which also has high UVA protection, and make sure you apply it generously and frequently when in the sun, preferably every 2 to 3 hours
- Sunscreens should not be used as an alternative to clothing or shade - rather they offer additional protection. No sunscreen will provide 100% protection
- Do not use sun beds
- Check your skin for changes once a month. A friend or family member can help you with this particularly with checking your back. If there is a new or changing lump or skin lesion, if it has been present for more than 6 weeks, is getting bigger, scabs or bleeds go to your doctor and have it looked at.

3) Percutaneous Liver Biopsy

What is a liver biopsy?

In a liver biopsy, the doctor will take a very small piece of your liver (about 1/50,000th of your liver) to send for further tests. Most liver biopsies are examined under the microscope by a pathologist. Sometimes, a piece of the liver is sent to microbiology to see if there is any infection in the liver. Your doctor will explain to you why he/she thinks you need a liver biopsy. A liver biopsy may be done to look for the cause of liver abnormalities, to assess if, and how much, the liver is damaged and/or to help in planning treatment.

Are there any alternatives to a liver biopsy?

There is no other procedure that will give your doctor the same information as a liver biopsy. A biopsy can be done through the side (percutaneous) or, if there are problems with the blood's ability to clot, through a long needle inserted through a small cut in the neck and, under X-ray guidance, passed into the liver.

How is the liver biopsy done?

You will need to be in hospital for at least 4-6 hours after the biopsy. This may require admission either overnight or may be done as a day case. Before the liver biopsy is done, the doctor or nurse will check your blood to make sure that clotting is within acceptable limits. Generally you will be seen in the pre-screening clinic and have your bloods taken, MRSA screen and complete a health screening questionnaire. If you are taking certain medicines, drugs or tablets that may affect bleeding and clotting you may be asked to stop them for a few days or a week or two before the biopsy. The pre-screening nurse may telephone you to check your drugs and tablets with you before you come to hospital.

Sometimes, perhaps because of liver disease, the blood does not clot as well as is needed and you may be given some blood factors (plasma) to help your blood clot. Sometimes you may also need platelets to be given. If it is unsafe the doctor will not proceed with the biopsy.

If you are nervous, do ask the doctor about having sedation before the biopsy. Please let us know about this as early as possible because injected sedation is not always available. In order to give injected sedation, an extra nurse or doctor needs to be present to give you the injection and monitor you during the procedure therefore telling us in advance will help us plan for this. You may also be able to take a 'relaxing' tablet by mouth before the procedure, if we know to prescribe it in advance.

You will be asked to give written consent to the biopsy before it is carried out. Please do not hesitate to ask the doctor if you have any questions at all.

Most biopsies are now done with ultrasound so you will be taken to the Imaging department on your bed or trolley. During the biopsy, you will have to lie flat on your bed or trolley. The ultrasound will be used to show the doctor exactly where your liver is and the best place for the biopsy. The liver is normally situated on your right side just under the lower ribs. Biopsy is often taken here or in the front at the top of your tummy. The doctor will explain the procedure to you before doing the actual biopsy.

The doctor will put some disinfectant on your skin and then inject a local anaesthetic. This may sting before the skin goes numb. You may also feel the local anaesthetic deeper inside before it works. It takes a few minutes to have its full effect. The doctor will then take the biopsy. This usually takes only a few seconds. You will be asked to hold your breath for a few seconds while it is being done. Sometimes we need to take 2 or 3 samples to be sure to have enough to analyse.

When the biopsy is done, a plaster will be put on your skin at the biopsy spot. You will have to lie on the bed or trolley for at least 6 hours. During that time, the nurse will take your blood pressure and pulse. This will be every 15 minutes for the first 2 hours, then less frequently. You will be able to eat and drink a little during this time.

Will it hurt?

The procedure may be uncomfortable for you and there may be some pain either at the site of the biopsy or in your right shoulder. Normally, the pain is controlled by local anaesthetic but sometimes you might need to take tablets to control the pain. The doctor will write you a prescription for this. If you are in pain, please do ask the nurse for pain relief.

What are the complications?

The vast majority of liver biopsies are done without any complication. The most common complication is pain and this can usually be controlled with pain killers.

Bleeding is a potential problem, and for this reason the doctor will check that your blood clots normally and the nurse will take your blood pressure and pulse after the procedure. Bleeding may affect 1 in 200 biopsies. Often this settles by itself, but we know that 1 in 4 patients who bleed may require a blood transfusion. Very rarely (perhaps less than 1 in 500 of the patients who have this problem) an operation is required to stop the bleeding.

As the doctor is unable to see the whole liver when performing a biopsy, there is the possibility that the needle might injure other organs, such as the gall bladder or, very rarely, the kidney or bowel. This happens in less than 1 in 1000 cases. Sometimes, needle injury to other organs does cause problems, however very severe complications are extremely rare.

After the liver biopsy

After the liver biopsy, you will need to stay in your bed for 4-6 hours, then if everything is satisfactory you may be free to go home. Please arrange for someone to collect you on discharge rather than drive yourself. You must have a responsible adult with you on the first night following your biopsy and be within a 30 minute drive from a hospital. If you have received sedation you should not drive or operate machinery for 36 hours.

If you have any discomfort you can usually take paracetamol, but if the pain is severe, let the hospital know.

Please ask about what arrangements will be made for you to be informed of the biopsy report.

It is important that you do not do anything strenuous within 36 hours after a liver biopsy. If you have any pain or any other untoward effect, which may be related to the liver biopsy please ring one of the contact numbers.

Appendix B: Multiple-choice comprehension questions used in Pilot 1

1) Femoral Hernia

1) Why might it be difficult to locate the hernia using the Laparoscopic technique?

Answer: The small incisions limit what the surgeon can see during surgery

Literal: Sometimes it may be difficult for the surgeon to see or handle tissue safely

Misconception: A small hernia may be difficult for the surgeon to locate

Near-miss inference: The technique is difficult so requires an expert surgeon to be able to locate the hernia

2) Why might treatment for a femoral hernia reduce the chance of it reoccurring?

Answer: If a mesh is used during treatment this will reinforce the abdominal canal

Literal: The weakness in the abdominal canal is repaired during treatment

Misconception: The section of herniated tissue is returned to the correct place after treatment

Near-miss inference: If the surgeon is able to locate the hernia during surgery it will be effectively repaired

3) Why might being a long-term smoker affect the likelihood of developing a femoral hernia?

Answer: A cough developed through long-term smoking will increase the pressure inside the abdomen

Literal: Activities including coughing increase the likelihood of developing a hernia

Misconception: Long-term smokers are more likely to cough more, causing excessive straining

Near-miss inference: Long-term smokers typically have poor health, meaning a hernia is more likely

4) Why might immediate surgery to treat a strangulated femoral hernia be so important?

Answer: Without immediate treatment, the herniated tissue will rapidly die

Literal: If left untreated the hernia can cause severe medical problems and complications

Misconception: A strangulated hernia blocks blood flow which is a medical emergency

Near-miss inference: Rapid treatment of a strangulated hernia provides a better recovery with less risk of complications

5) Harry has a developed small lump in his groin. Since it's not painful or causing any problems, Harry doesn't think he should see a doctor. Why is Harry incorrect?

Answer: It could be a femoral hernia, so it is important to get medical advice as soon as possible, as it could get worse

Literal: Even if there are no troublesome symptoms, it could be a femoral hernia, which causes a grape-sized lump in the groin

Misconception: The lump could be anything - it is important to speak to a doctor for advice and a diagnosis of the problem

Near-miss inference: Seeking medical advice early is important, as treatments for many conditions are most effective at an early stage

2) Basal Cell Carcinoma

1) Why might basal cell carcinomas be less likely to occur on the feet?

Answer: The feet are more likely to be covered up and less exposed to the sun, so basal cell carinomas are less likely to occur there

Literal: Basal cell carcinomas are more likely to occur on the face, scalp, neck, back and chest

Misconception: It is quite unlikely for the feet to be sunburnt, so basal cell carcinomas are less common there

Near-miss inference: The use of adequate sun protection on your feet will reduce the chance of a basal cell carcinoma developing

2) If a safety margin of skin was not removed during surgery to remove the basal cell carcinoma, what might happen?

Answer: The basal cell carcinoma could grow back after surgery

Literal: Not all of the roots of the basal cell carcinoma may have been removed

Misconception: The basal cell carcinoma could spread to other areas

Near-miss inference: Further treatment may be needed which may be more difficult

3) Why might surgery not be appropriate for some people?

Answer: If people have neglected lesions that have grown into nearby structures, other forms of treatment may be more appropriate

Literal: Radiotherapy may be more appropriate for some people, which destroys the cancer cells while doing as little harm as possible to normal cells

Misconception: A number of factors, including as age and overall health, inform the appropriateness of surgery

Near-miss inference: Surgery can be particularly difficult where the lesion is large or on an area which is problematic to access

4) Why might a skin graft sometimes be needed after surgery to remove a basal cell carcinoma?

Answer: For large or deep basal cell carcinomas, stitches aren't adequate to close the wound

Literal: After surgery, the area is usually stitched together though sometimes a skin graft is needed

Misconception: Surgery will leave an open wound in the skin and a skin graft is an effective way of covering it

Near-miss inference: The removal of the safety margin of skin may mean that sometimes stitches are not appropriate

5) Jenny thinks she might have a basal cell carcinoma but has never been sunburnt and protects her skin from damage. Why might Jenny be correct?

Answer: Exposure to even small amounts of sun can lead to basal cell carcinomas, even if protection from the sun is regularly used

Literal: Basal cell carcinomas arise due to too much sun exposure in people with fair skin

Misconception: Any exposure to the sun, throughout your life, causes damage to the skin and increases the likelihood of developing skin cancer

Near-miss inference: Skin damage caused by UV exposure may have occurred indirectly through the use of a sunbed

3) Percutaneous Liver Biopsy

1) How might the unavailability of an ultrasound scanner affect the liver biopsy procedure?

Answer: There is an increased risk of damage to other organs, as the doctor may be less accurate when collecting the biopsy

Literal: Without an ultrasound scanner, the doctor will not know exactly where your liver is and the best place for the biopsy

Misconception: If the liver is not able to be seen, a biopsy may not be able to be taken, so the procedure will not be able to be carried out

Near-miss inference: The unavailability of an ultrasound scanner will not affect the alternative method of taking the biopsy - using a needle inserted into the neck

2) Why is it important to collect a sufficient number of samples during the procedure?

Answer: To ensure that testing can be carried out accurately to give a diagnostic result

Literal: During the procedure, 2 or 3 samples need to be taken to be sure to have enough to analyse

Misconception: To allow different areas of the liver to be tested to determine the location of the problem

Near-miss inference: Collecting an insufficient number of samples during the procedure may mean that it will have to be repeated

3) Why might you be asked to hold your breath when the biopsy is taken?

Answer: Holding your breath will help to prevent damaging other organs while the sample is taken

Literal: While the sample is taken, it is important to hold your breath for a few seconds

Misconception: So that there will be no movement of your lungs and chest wall during the biopsy

Near-miss inference: Expanding the chest cavity, by holding your breath, allows the sample to be collected more accurately

4) Following a liver biopsy, why might it not be advised to take aspirin to manage pain?

Answer: To ensure that blood is clotting effectively post-procedure, aspirin should be avoided after the biopsy

Literal: Aspirin must not be taken a week (7 days) prior to the liver biopsy, as it can result in complications

Misconception: Painkillers are administered throughout the procedure, taking more could lead to an overdose

Near-miss inference: After any surgical treatment, it is not advisable to self-prescribe forms of pain relief - medical advice should be sought

5) Fatima needs a liver biopsy and has been advised by her doctor that the procedure will need to be carried out through her neck. Why might this be the case?

Answer: When blood is not clotting normally, this approach will mitigate the potential risk of bleeding out

Literal: A needle inserted into the neck is used if there are problems with the blood's ability to clot

Misconception: The needle travels along a vein which is less invasive and damaging, reducing the likelihood of bleeding

Near-miss inference: This approach is safer and more accurate, as the biopsy sample is collected using X-ray guidance

Appendix C: QRI multiple-choice responses

As material from the QRI cannot be reproduced, the question stems and correct answers are not provided here. Where the type of distractor is shown strikethrough, these items do not closely correspond to the type shown, for reasons discussed in 3.2.1 Pilot 1: Materials, but approximate the classification to a varying extent.

'Life Cycle of Stars' Passage 1

Question 1

Literal: The life cycle of stars

Misconception: The composition of space between stars

Near miss-inference: Where stars are made

Question 2

Literal: 10 million years

~~Misconception~~: 1 billion years

~~Near-miss inference~~: 1 million years

Question 3

Literal: Where stars are born

Misconception: The space between stars

Near-miss inference: A clump of hydrogen particles

Question 4

Literal: As it travels through space, it will collect dust and gas

Misconception: It will get larger so more able to catch dust

Near-miss inference: It will encounter more dust in other parts of space

Question 5

Literal: A star that doesn't yet shine ordinary light

Misconception: A newly-formed, baby star

Near-miss inference: Occurs when nebulae are sufficiently large

Question 6

Literal: Scientists identify protostars within nebulae using infrared telescopes

Misconception: Due to the movement of the gas and dust forming into a whole

Near-miss inference: Scientists can infer protostars from the effect of gravity on surrounding objects

Question 7

Literal: When nuclear fusion produces great amounts of energy, a star comes to life

Misconception: The pressure in the core of the protostar becomes extremely high

Near-miss inference: When the protostar begins to give out ordinary light, not just infrared light

Question 8

Literal: Its density

Misconception: Its temperature

Near-miss inference: Its size

Question 9

Literal: Gravity causes matter to be attracted to other matter

Misconception: Gravity holds the dust and gas in the nebula together

Near-miss inference: Gravity increases the pressure of matter packed tightly together

Question 10

Literal: nuclear fusion

~~Misconception~~: hydrogen

~~Near-miss inference~~: nuclear fission

‘Life Cycle of Stars’ Passage 2

Question 11

Literal: Low and high-mass stars

Misconception: How stars collapse and die

Near-miss inference: Different types of stars

Question 12

Literal: Gravity

Misconception: Hydrogen

Near-miss inference: Pressure

Question 13

Literal: After the supply of hydrogen in the core has run out, there is no longer material for nuclear reactions to occur, so the star becomes a red giant

Misconception: The red giant stage begins after the main-sequence period has ended and the hydrogen is exhausted

Near-miss inference: Changes in pressure first cause the core to collapse, the star then grows as the temperature then increases

Question 14

Literal: 50 million years

~~Misconception~~: 50 billion years

~~Near-miss inference~~: 500 billion years

Question 15

Literal: The size of Mars

~~Misconception~~: The size of the Moon

Near-miss inference: The size of a planet

Question 16

Literal: Neutron star stage

Misconception: Supernova stage

~~Near-miss inference~~: Late high-mass star stage

Question 17

Literal: The expansion of the outer layers

Misconception: Extreme pressure

Near-miss inference: Changes within the core

Question 18

Literal: A black dwarf

Misconception: A white dwarf

~~Near-miss inference~~: A red giant

Question 19

Literal: Nothing can escape from it

Misconception: It is too small

Near-miss inference: There is no nuclear fusion

Question 20

Literal: Scientists have no real proof so far

Misconception: It is not possible to see a black hole

Near-miss inference: They do not emit heat so infrared telescopes cannot see them

'American Immigration' Passage 1

Question 1

Answer: Reasons why immigrants came to America

Literal: Push and pull factors

Misconception: Immigration in the late 1800s to early 1900s

Near-miss inference: Where American immigrants came from

Question 2

Literal: Farm machines replaced farm workers

Misconception: To escape poverty

Near-miss inference: The European populations had grown too large

Question 3

Literal: As European populations grew, land for farming became scarce

Misconception: Farms could not produce enough food for larger populations

Near-miss inference: The amount of land is fixed, so no more land could be made available

Question 4

Literal: Persecution and emigrant relatives

~~Misconception~~: Promise of freedom and hardship

~~Near-miss inference~~: Emigrant relatives and promise of freedom

Question 5

Literal: Hardship

~~Misconception~~: Poverty

~~Near-miss inference~~: Promise of freedom

Question 6

Literal: He was likely to be a bold family member

Misconception: He would find it easier to get a job

Near-miss inference: He would prepare for the rest of the family to emigrate

Question 7

Literal: Persecution and emigrant relatives

~~Misconception~~: Hardship and persecution

~~Near-miss inference~~: Promise of freedom and hardship

Question 8

Literal: Once settled, the newcomers helped pull neighbours to America

Misconception: Jealousy over the neighbour's better life in America

Near-miss inference: Settled neighbours helped build communities

Question 9

Literal: Jobs were a pull factor for immigrants

Misconception: European workers wanted to move for a better future

Near-miss inference: There were not enough workers in America

Question 10

Literal: Railroads posted notices in Europe advertising cheap land

Misconception: To attract people to live in the West of America

Near-miss inference: Europeans needed land to support their families

'American Immigration' Passage 2

Question 11

Literal: The voyage to America and becoming American

Misconception: Immigration in the late 1800s to early 1900s

Near-miss inference: What life was like for new immigrants

Question 12

Literal: Leaving home required great courage

Misconception: The journey was very long

Near-miss inference: Ship owners were cruel and greedy

Question 13

Literal: Immigrants travelled on the cheapest berths

Misconception: There was no medicine available

Near-miss inference: Steerage was not regularly cleaned

Question 14

Literal: Immigrants immediately set out to find work

~~Misconception~~: Advertisements in the local area

Near-miss inference: Overseas factory recruitment initiatives

Question 15

Literal: Previously they had little need for money

~~Misconception~~: To send money to their family overseas

Near-miss inference: They needed it to improve their lives

Question 16

Literal: Most immigrants stayed where they landed

Misconception: It was better than where they had come from

Near-miss inference: It was difficult to find employment elsewhere

Question 17

Literal: The process of settling into another country

Misconception: The process of taking on an American identity

~~Near-miss inference~~: The process of understanding the beliefs and ideas of others

Question 18

Literal: Religion

~~Misconception~~: Ethnicity

~~Near-miss inference~~: Location

Question 19

Literal: Many Americans opposed the increase in immigration

Misconception: Americans thought that the immigrants would take the jobs

Near-miss inference: They thought that there were too many immigrants

Question 20

Literal: Children learned English in school

Misconception: Parents did not want to assimilate

Near-miss inference: Parents could not afford to buy children American-style clothes

Appendix D: HLVA items and multiple-choice responses

1) Intravenous

Answer: administered into a vein

Semantic: someone that often intervenes

Misconception: injection of a substance

Insufficient: entering through the skin

2) Biopsy

Answer: the removal of tissue for analysis

Semantic: performed on bodies after death

Misconception: taking a sample of cells

Insufficient: determine what something is made of

3) Antibiotic

Answer: a drug to kill bacteria

Semantic: fights and destroys living cells

Misconception: medication to treat infection

Insufficient: a substance to help the body

4) Radiotherapy

Answer: treat disease using energy

Semantic: treatment to improve a condition

Misconception: burns away bad cells

Insufficient: target and treat cancer diseases

5) Prescription

Answer: instructions for drugs you need

Semantic: being told what you must do

Misconception: taken to and dispensed by the pharmacist

Insufficient: list of various medications and amounts

6) Oral

Answer: when taken by mouth

Semantic: when speaking or verbalising

Misconception: another word for mouth

Insufficient: relating to part the body

7) Injection

Answer: inserting a needle into the body

Semantic: giving a small amount of a substance

Misconception: inserting a needle into a vein

Insufficient: using a needle to administer medication

8) Anaesthetic

Answer: a substance which reduces sensation

Semantic: a substance which abolishes visual experience

Misconception: a drug which puts you to sleep

Insufficient: a drug given before you have surgery

9) Symptoms

Answer: the signs of a disease

Semantic: side effects of drugs

Misconception: when you feel unwell

Insufficient: departure from normal feelings

10) Medication

Answer: given to improve a condition

Semantic: given to cure disease

Misconception: given to make you well

Insufficient: given to improve health

11) Infection

Answer: an illness caused by pathogens

Semantic: a contagious condition

Misconception: a wound which has become contaminated

Insufficient: an invasion of the body

12) Scan

Answer: to visually examine interior parts of the body

Semantic: to analyse an area to detect features of interest

Misconception: to see the structures of the body more closely

Insufficient: using computer-assisted techniques, such as ultrasonography

13) Liver

Answer: an organ involved in metabolising and detoxifying

Semantic: respiration takes place within this organ

Misconception: an organ which regulates blood supply

Insufficient: the largest internal organ located in the abdomen

14) Stroke

Answer: an interruption to the blood flow to an area of the brain

Semantic: an emergency condition that requires prompt treatment

Misconception: loss of speech and function on one side of the body

Insufficient: caused by a blood clot blocking the supply of oxygen

15) Ward

Answer: a room in a hospital with beds for patients

Semantic: a patient who is under the care of medical practitioner

Misconception: department-specific rooms where patients receive treatment

Insufficient: an area where patients recover and recuperate

16) Cancer

Answer: illness caused by abnormal multiplication of cells

Semantic: a disease which is hard to eradicate

Misconception: an internal or external growth of the body

Insufficient: a common illness which can often be fatal

Appendix E: Stimulus texts used in Pilot 2 and Study 1

1) Aspergillosis

Aspergillosis is a condition caused by aspergillus mould. There are several different types of aspergillosis. Most affect the lungs and cause breathing difficulties.

How you get aspergillosis

Aspergillosis is usually caused by inhaling tiny bits of mould. The mould is found in lots of places, including: soil, compost and rotting leaves, plants, trees and crops, dust, damp buildings, and air conditioning systems.

You can't catch aspergillosis from someone else or from animals.

Most people who breathe in the mould don't get ill.

Aspergillosis is rare in healthy people.

You're usually only at risk of aspergillosis if you have: a lung condition – such as asthma, cystic fibrosis or chronic obstructive pulmonary disease (COPD), a weakened immune system – for example, if you have had an organ transplant or are having chemotherapy or had tuberculosis (TB) in the past.

Symptoms of aspergillosis

Symptoms of aspergillosis include: shortness of breath, a cough – you may cough up blood or lumps of mucus, wheezing (a whistling sound when breathing), a high temperature of 38C or above, and weight loss.

If you already have a lung condition, your existing symptoms may get worse.

See a GP if you have: a cough for more than 3 weeks, a lung condition that's getting worse or harder to control with your usual treatment, a weakened immune system and symptoms of aspergillosis.

Get an urgent GP appointment if you cough up blood. Call 111 if you can't see your GP.

What happens at your appointment

Your GP will check for an obvious cause of your symptoms, like a chest infection or asthma. If they're not sure what the problem is, they may refer you to a specialist for tests such as: X-rays and scans, blood tests or tests on a sample of mucus, allergy tests, or a bronchoscopy – where a thin, flexible tube with a camera at the end is used to look in your lungs.

2) Autosomal dominant polycystic kidney disease

Autosomal dominant polycystic kidney disease (ADPKD) is an inherited condition that causes small, fluid-filled sacs called cysts to develop in the kidneys.

Although children affected by ADPKD are born with the condition, it rarely causes any noticeable problems until the cysts grow large enough to affect the kidneys' functions.

In most cases, this doesn't occur until a person is between 30 and 60 years of age. Less commonly, children or older people may have noticeable symptoms as a result of ADPKD.

When ADPKD reaches this stage, it can cause a wide range of problems, including: abdominal (tummy) pain, high blood pressure (hypertension), blood in the urine (haematuria) – which may not always be noticeable to the naked eye, potentially serious upper urinary tract infections (UTIs), and kidney stones.

Kidney function will gradually deteriorate until so much is lost that kidney failure occurs.

What causes ADPKD?

ADPKD is caused by a genetic fault that disrupts the normal development of some of the cells in the kidneys and causes cysts to grow.

Faults in one of two different genes are known to cause ADPKD. The affected genes are: PKD1 – which accounts for 85% of cases , and PKD2 – which accounts for 15% of cases.

Both types of ADPKD have the same symptoms, but they tend to be more severe in PKD1.

A child has a one in two (50%) chance of developing ADPKD if one of their parents has the faulty PKD1 or PKD2 gene.

Autosomal recessive polycystic kidney disease (ARPKD) is a rarer type of kidney disease which can only be inherited if both parents carry the faulty gene and in this type problems usually start much earlier, during childhood.

3) Non-melanoma skin cancer

Skin cancer is one of the most common cancers in the world. Non-melanoma skin cancer refers to a group of cancers that slowly develop in the upper layers of the skin.

The term non-melanoma distinguishes these more common types of skin cancer from the less common skin cancer known as melanoma, which can be more serious.

In the UK, more than 100,000 new cases of non-melanoma skin cancer are diagnosed each year. It affects more men than women and is more common in the elderly.

Symptoms of non-melanoma cancer

The first sign of non-melanoma skin cancer is usually the appearance of a lump or discoloured patch on the skin that continues to persist after a few weeks, and slowly progresses over months or sometimes years. This is the cancer, or tumour.

In most cases, cancerous lumps are red and firm and sometimes turn into ulcers, while cancerous patches are usually flat and scaly.

Non-melanoma skin cancer most often develops on areas of skin regularly exposed to the sun, such as the face, ears, hands, shoulders, upper chest and back.

When to get medical advice

See your GP if you have any skin abnormality, such as a lump, ulcer, lesion or skin discolouration that hasn't healed after four weeks. While it's unlikely to be skin cancer, it's best to be sure.

Types of non-melanoma skin cancer

Non-melanoma skin cancers usually develop in the outermost layer of skin (epidermis), and are often named after the type of skin cell from which they develop.

The two most common types of non-melanoma skin cancer are:

basal cell carcinoma (BCC) – also known as a rodent ulcer, BCC starts in the cells lining the bottom of the epidermis and accounts for about 75% of skin cancers

squamous cell carcinoma (SCC) – starts in the cells lining the top of the epidermis and accounts for about 20% of skin cancers

4) Behcet's disease

Behcet's disease, or Behcet's syndrome, is a rare and poorly understood condition that results in inflammation of the blood vessels and tissues.

Confirming a diagnosis of Behcet's disease can be difficult because the symptoms are so wide-ranging and general (they can be shared with a number of other conditions).

Symptoms of Behcet's disease

The main symptoms of Behcet's disease include: mouth ulcers, red, painful eyes and blurred vision, acne-like spots, headaches, and painful, stiff and swollen joints.

In severe cases, there's also a risk of serious and potentially life-threatening problems, such as permanent vision loss and strokes.

Most people with the condition experience episodes where their symptoms are severe (flare-ups or relapses), followed by periods where the symptoms disappear (remission).

Over time, some of the symptoms can settle down and become less troublesome, although they may never resolve completely.

Diagnosing Behcet's disease

There's no definitive test that can be used to diagnose Behcet's disease.

Several tests may be necessary to check for signs of the condition, or to help rule out other causes, including: blood tests, urine tests, scans, such as X-rays, a computerised tomography (CT) scan or a magnetic resonance imaging (MRI) scan, a skin biopsy, or a pathergy test – which involves pricking your skin with a needle to see if a particular red spot appears within the next day or two; people with Behcet's disease often have particularly sensitive skin

Current guidelines state a diagnosis of Behcet's disease can usually be confidently made if you've experienced at least 3 episodes of mouth ulcers over the past 12 months and you have at least 2 of the following symptoms: eye inflammation, skin lesions (any unusual growths or abnormalities that develop on the skin), or pathergy (hypersensitive skin).

Other potential causes also need to be ruled out before the diagnosis is made.

5) Schistosomiasis

Schistosomiasis, also known as bilharzia, is an infection caused by a parasitic worm that lives in fresh water in subtropical and tropical regions.

The parasite is most commonly found throughout Africa, but also lives in parts of South America, the Caribbean, the Middle East and Asia.

You often don't have any symptoms when you first become infected with schistosomiasis, but the parasite can remain in the body for many years and cause damage to organs such as the bladder, kidneys and liver.

The infection can be easily treated with a short course of medicine, so see your GP if you think you might have it.

How you get schistosomiasis

The worms that cause schistosomiasis live in fresh water, such as: ponds, lakes, rivers, reservoirs, and canals.

Showers that take unfiltered water directly from lakes or rivers may also spread the infection, but the worms aren't found in the sea, chlorinated swimming pools or properly treated water supplies.

You can become infected if you come into contact with contaminated water – for example, when paddling, swimming or washing – and the tiny worms burrow into your skin. Once in your body, the worms move through your blood to areas such as the liver and bowel.

After a few weeks, the worms start to lay eggs. Some eggs remain inside the body and are attacked by the immune system, while some are passed out. Without treatment, the worms can keep laying eggs for several years.

If the eggs pass out of the body into water, they release tiny larvae that need to grow inside freshwater snails for a few weeks before they're able to infect another person. This means it's not possible to catch the infection from someone else who has it.

6) Bornholm Disease

Bornholm disease (also called pleurodynia) is a viral infection that causes pain in the chest or upper tummy and flu-like symptoms.

It usually clears up by itself after a few days, but can sometimes last longer (up to 3 weeks).

Bornholm disease mainly affects children and young adults.

Symptoms of Bornholm disease

The main symptom of Bornholm disease is a severe, stabbing chest pain, which is often worse when you breathe deeply, cough or move.

The pain tends to come and go, with episodes lasting 15 to 30 minutes.

In very severe cases, the pain can make it difficult to breathe and the affected area may be tender.

Other symptoms of Bornholm disease include: tummy pain, high temperature (fever), headache, sore throat, cough, and aching muscles.

These symptoms usually start suddenly and last for a few days. They can sometimes last longer (up to 3 weeks), or they can come and go for a few weeks before eventually clearing up.

When to get medical help

If you have chest pain, it's important to get it checked out, particularly if it's severe and comes on suddenly.

Bornholm disease can be serious for newborn babies, so if you're in the late stages of pregnancy or have a newborn baby and you've come into contact with someone with the condition, ask your midwife or GP for advice.

Treating Bornholm disease

There's no specific treatment for Bornholm disease. The infection usually clears up on its own within a week.

As the condition is caused by a virus, it can't be treated with antibiotics. You can use over-the-counter painkillers, such as paracetamol and ibuprofen, to help with any pain.

Newborn babies at risk of getting Bornholm disease may be treated with immunoglobin to make the effects of the virus less severe and help prevent complications.

7) Brugada syndrome

Brugada syndrome is a rare but serious condition that affects the way electrical signals pass through the heart. It can cause the heart to beat dangerously fast.

These unusually fast heartbeats – known as an arrhythmia – can be life threatening in some cases.

Brugada syndrome is usually caused by a faulty gene that's inherited by a child from a parent. A simple heart test can be done to see if you have it.

Symptoms of Brugada syndrome

Many people with Brugada syndrome don't have any symptoms and don't realise they have it. Some people experience: blackouts, fits (seizures), occasional noticeable heartbeats (palpitations), chest pain, breathlessness, or dizziness.

These symptoms can occur at any time, but are sometimes triggered by something such as a high temperature (fever), drinking lots of alcohol, or dehydration.

Symptoms typically first appear at around 30-40 years of age, but they can occur at any age. They're more common in men than women or children.

When to get medical advice

See your GP if: you have unexplained blackouts or seizures, one of your parents, siblings or children has been diagnosed with Brugada syndrome – this may mean you're also at risk, or a close family member has died suddenly with no explanation – this can sometimes be the result of an undiagnosed heart problem like Brugada syndrome.

They can refer you to a specialist heart doctor for some simple tests to check if you have Brugada syndrome or any other heart problem.

If you've already been diagnosed with Brugada syndrome, contact your specialist as soon as possible if you experience any symptoms.

Tests for Brugada syndrome

The main test for Brugada syndrome is a test of the heart's electrical activity, known as an electrocardiogram (ECG). This is usually done in hospital.

During an ECG, small sensors are attached to your arms, legs and chest. These are connected to a machine that measures the electrical signals produced by your heart each time it beats.

8) Transient ischaemic attack (TIA)

A transient ischaemic attack (TIA) or "mini stroke" is caused by a temporary disruption in the blood supply to part of the brain.

The disruption in blood supply results in a lack of oxygen to the brain. This can cause sudden symptoms similar to a stroke, such as speech and visual disturbance, and numbness or weakness in the face, arms and legs.

However, a TIA doesn't last as long as a stroke. The effects often only last for a few minutes or hours and fully resolve within 24 hours.

Symptoms of a TIA

The main symptoms of a TIA can be remembered with the word FAST: Face-Arms-Speech-Time.

Face – the face may have dropped on one side, the person may not be able to smile, or their mouth or eye may have dropped.

Arms – the person with suspected stroke may not be able to lift both arms and keep them there because of arm weakness or numbness in one arm.

Speech – their speech may be slurred or garbled, or the person may not be able to talk at all, despite appearing to be awake.

Time – it is time to dial 999 immediately if you see any of these signs or symptoms.

When to seek medical advice

In the early stages of a TIA, it's not possible to tell whether you're having a TIA or a full stroke, so it's important to phone 999 immediately and ask for an ambulance.

Even if the symptoms disappear while you're waiting for the ambulance to arrive, an assessment in hospital should still be carried out.

A TIA is a warning that you may be at risk of having a full stroke in the near future, and an assessment can help doctors to determine the best way to reduce the chances of this happening.

9) Cholesteatoma

A cholesteatoma is an abnormal collection of skin cells deep inside your ear.

They're rare but, if left untreated, they can damage the delicate structures inside your ear that are essential for hearing and balance.

A cholesteatoma can also lead to: an ear infection – causing discharge from the ear, hearing loss – this can be permanent, vertigo – the sensation that you, or the world around you, is spinning, tinnitus – hearing sounds coming from inside the body, rather than from an outside source, and damage to your facial nerve – this can cause weakness in half your face.

In very rare cases, an infection can spread into the inner ear and brain, leading to a brain abscess or meningitis.

Symptoms of cholesteatoma

A cholesteatoma usually only affects one ear. The two most common symptoms are: a persistent or recurring watery, often smelly, discharge from the ear, which can come and go or may be continuous, and a gradual loss of hearing in the affected ear.

Some people may experience slight discomfort in their ear.

When to see your GP

See your GP if you have problems with your hearing or a watery discharge from your ear.

Your GP may examine your ear with an otoscope – an instrument with a light and magnifying glass.

They may suspect a cholesteatoma from your symptoms, but it can be difficult to confirm because a build-up of pus inside the ear often blocks it from view.

If your GP thinks your symptoms could just be an ear infection, they may offer you treatment for this first and ask to see you again once you've completed it.

If they think you have a cholesteatoma, they should refer you to an ear, nose and throat (ENT) specialist for further tests.

10) Isovaleric acidaemia

Isovaleric acidaemia (IVA) is a rare, but potentially serious, inherited condition. It means the body can't process the amino acid leucine (amino acids are "building blocks" of protein). This causes a harmful build-up of the substance in the blood and urine.

Normally, our bodies break down protein foods like meat and fish into amino acids. Any amino acids that aren't needed are usually broken down and removed from the body.

Babies with IVA are unable to fully break down the amino acid leucine.

Normally, leucine is broken down into a substance called isovaleric acid, which is then converted into energy. Babies with IVA don't have the enzyme that breaks down isovaleric acid, leading to a harmfully high level of this substance in the body.

Diagnosing IVA

At around 5 days old, babies are now offered newborn blood spot screening to check if they have IVA. This involves pricking your baby's heel to collect drops of blood to test.

If IVA is diagnosed, treatment can be given straight away to reduce the risk of serious complications. Treatment includes a special diet, advice and sometimes medication.

With early diagnosis and the correct treatment, the majority of children with IVA are able to live healthy lives. However, treatment for IVA must be continued for life.

Without treatment, severe and life-threatening symptoms can develop in some children, including seizures (fits) or falling into a coma. Some children with untreated IVA are also at risk of brain damage and developmental delay.

Symptoms of IVA

The symptoms of IVA aren't the same for everyone with the condition and some people may have more severe or frequent symptoms.

Symptoms sometimes appear within the first few days or weeks after birth and may include: developing a distinctive odour of "sweaty feet", poor feeding or loss of appetite, and weight loss.

Appendix F: Multiple-choice comprehension questions used in Pilot 2

Response options provided in parentheses, correct response is italicised.

1) Aspergillosis

1) How likely is an average person to suffer from aspergillosis?

   (*very unlikely*, somewhat unlikely, somewhat likely, very likely)

2) Aspergillosis is what type of infection?

   (viral, bacterial, *fungal*, protozoan)

3) Which profession might be most likely to be exposed to aspergillus?

   (vet, baker, *gardener*, banker)

4) What might the symptoms of aspergillosis be mistaken for?

   (heart attack, *pneumonia*, food poisoning, multiple sclerosis)

2) Autosomal dominant polycystic kidney disease

1) How is ADPKD developed?

   (from infected livestock, *from your parents*, through air-borne particulates, from contaminated water)

2) Why does it take many years for the symptoms of ADPKD to develop?

   (fluid is retained in adulthood, the kidneys take time to grow, it doesn't cause problems for children, *the cysts develop slowly*)

3) What might the symptoms of ADPKD be mistaken for?

   (*bladder infection*, Crohn's disease, dengue fever, conjunctivitis)

4) A person with severe long-term ADPKD may need

   (regular blood transfusions, chemotherapy, *an organ transplant*, gene editing treatment)

3) Non-melanoma skin cancer

1) Which individual would be most likely to develop a non-melanoma skin cancer?

   (a younger lady, an older lady, a younger male, *an older male*)

2) What might a non-melanoma skin cancer be mistaken for?

   (arthritis, Athlete's foot, *ring worm*, leukaemia,)

3) Why might non-melanoma skin cancers be considered less serious than other forms of skin cancer?

   (they grow on the surface of the skin so are easily detected, *they grow slowly so don't spread quickly*, they typically don't cause other symptoms so are less severe, they are easy to remove so more likely be cured)

4) Which profession might be most likely to develop a non-melanoma skin cancer?

   (pilot, lawyer, teacher, *builder*)

4) Behcet's Disease

1) How likely is an average person to suffer from Behcet's disease?

   (*very unlikely*, somewhat unlikely, somewhat likely, very likely)

2) Getting treatment for Behcet's disease may be problematic because it is…?

   (few doctors understand it, *it is hard to diagnose*, there are few treatments, the symptoms are varied)

3) How can you minimise your risk of developing Behcets disease?

   (avoid environmental triggers including alcohol and stress, eat a healthy balanced diet with appropriate amounts of exercise, nothing because the condition is inherited, *nothing because the cause is not known*)

4) How might you distinguish between suffering from mouth ulcers and having Behcets disease?

   (if the ulcers become infected, if the ulcers are much sorer than would be expected, *if* *the ulcers recur several times a year*, if the ulcers last long periods of time before resolving)

5) Schistosomiasis

1) A person with schistosomiasis could be described as…?

   (a carrier, *a host*, anaemic, infectious)

2) What hobby might put you at greater risk of developing schistosomiasis?

   (birdwatching, surfing, *travelling*, running)

3) Why might symptoms take time to develop?

   (it takes the body's immune system time to identify the worm, the reaction builds up overtime until noticeable symptoms occur, chemicals released by the worm supress the immune system, *symptoms develop only when eggs are laid*)

4) Why might untreated schistosomiasis be serious?

   (the infection will be transmitted to others, *it damages major internal structures*, it becomes harder to treat over time, it makes additional infections more likely)

6) Bornholm Disease

1) How is Bornholm disease developed?

   (from contaminated water, *from infected individuals*, from your parents, from infected livestock)

2) What might the symptoms of Bornholm disease be mistaken for?

   (anaphylactic shock, a stroke, a seizure, *a heart attack*)

3) What might you take to cure Bornholm disease?

   (antibiotics, painkillers, anti-inflammatories, *nothing*)

4) Which group of people should be particularly worried about the risk of Bornholm disease?

   (*expectant mothers*, adult females, post-menopausal females, elderly adults)

7) Brugada Syndrome

1) How is Brugada syndrome developed?

   (*from your parents*, from contaminated water, from infected livestock, from infected individuals)

2) What might some of the symptoms of Brugada syndrome be mistaken for?

   (ulcerative colitis, *low blood pressure*, asthma, eczema)

3) Why is Brugada syndrome considered serious?

   (the symptoms are easily triggered, *it can be fatal*, there are no treatments for the condition, the symptoms can cause long-term injury)

4) Which group of people should be wary of the potential of developing Brugada syndrome?

   (younger females, *middle-aged males*, elderly adults, adults with weakened immune systems)

8) Transient Ischaemic attack

1) Which of the following might increase the likelihood of a TIA?

   (vitamin deficiency, low blood sugar, fluid retention, *high cholesterol*)

2) Why are TIAs generally considered less serious than related conditions?

   (it is easily treated, it is not fatal, *the effects are temporary*, the condition is very rare)

3) What might the symptoms of TIA be mistaken for?

   (a migraine, *a stroke*, a coma, an allergic reaction)

4) What’s the most important thing to do if you suspect TIA?

   (start chest compressions, see your GP, clear the airways, *act quickly*)

9) Cholesteatoma

1) If you were suffering from cholesteatoma, you would expect to have what type of problems?

   (*auditory*, visual, circulatory, pulmonary)

2) What might the symptoms of cholesteatoma be mistaken for?

   (lymphoma, conjunctivitis, *an infection*, a blood clot,)

3) How is a cholesteatoma developed?

   (from poor diet without exercise, from air-borne particulates, *from irregular tissue growth*, from your parents)

4) How likely is an average person to suffer from a cholesteatoma?

   (*very unlikely*, somewhat unlikely, somewhat likely, very likely)

10) Isovaleric acidaemia

1) What foods should someone with isovaleric acidaemia avoid?

   (tomatoes, *chicken*, sweets, coffee)

2) How is isovaleric acidaemia developed?

   (from exposure to toxic substances, from irregular tissue growth, from infected livestock, *from your parents*)

3) When are you most likely to be diagnosed with isovaleric acidaemia?

   (*childhood*, adolescence, adulthood, late-adulthood)

4) Why might symptoms of isovaleric acidaemia take time to develop?

   (*it takes time for Leucine to build up in the body*, it takes the body’s immune system time to respond to Leucine, some people have a greater resistance to Leucine, some medications can suppress the activity of Leucine delaying symptoms)

Appendix G: Altered multiple-choice comprehension questions used in Study 1

2) Autosomal dominant polycystic kidney disease

1) What could you do to reduce your chance of developing ADPKD?

(maintain a healthy weight, *nothing at all*, avoid regularly drinking alcohol, avoid sources of infection)

6) Bornholm Disease

4) Which group of people should be particularly worried about the risk of Bornholm disease?

(children, young adults, *expectant mothers*, all mothers)

7) Brugada Syndrome

1) What treatment might you receive to cure Brugada syndrome?

(*nothing*, antibiotics, antivirals, chemotherapy)

8) Transient Ischaemic attack

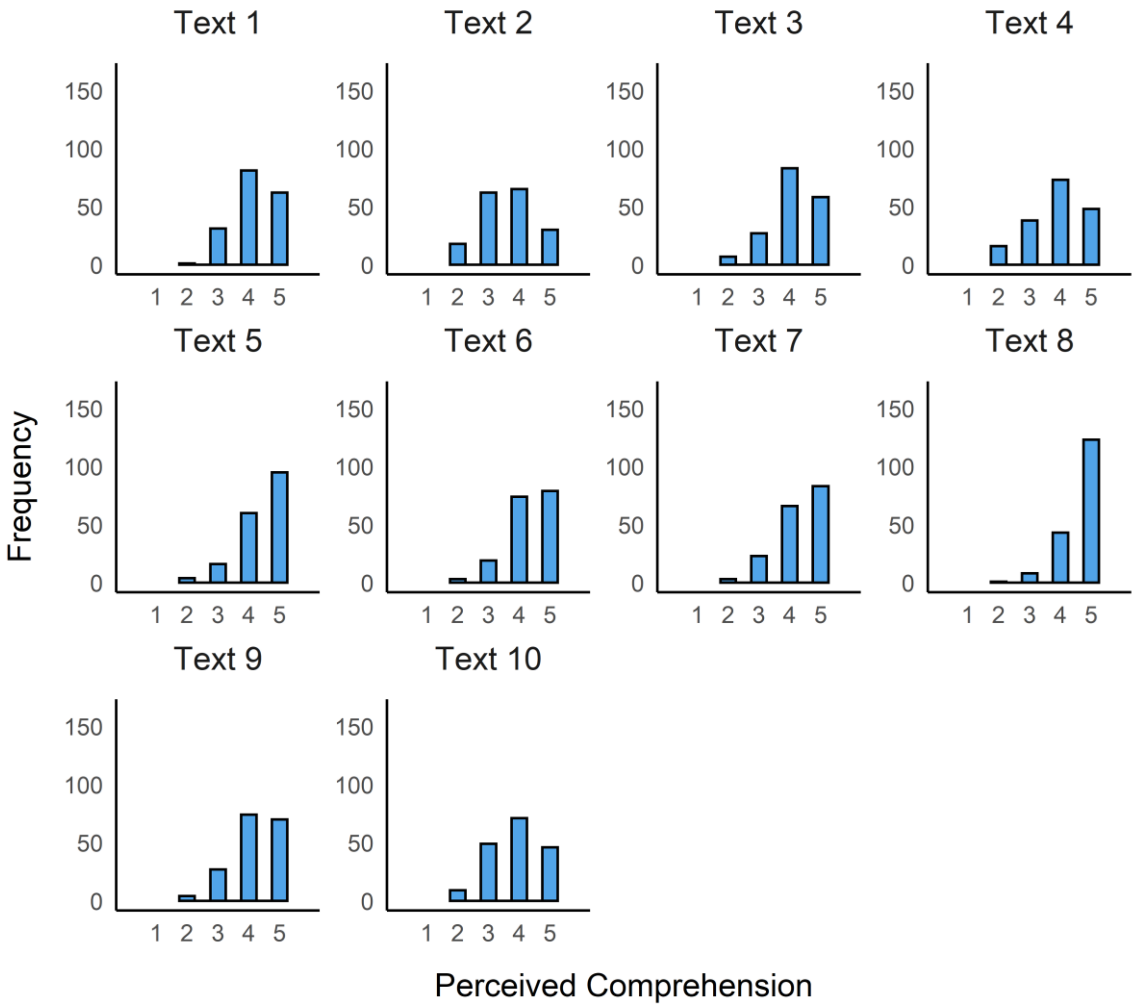
3) Why is it important to see a doctor after a TIA?

(to determine what caused the TIA, *you might be likely to have a stroke in the future*, to check that blood flow has returned to normal, you may need rehabilitation to reverse any damage)

Appendix H: Supplementary figure showing judgements by text in Study 1

Figure H.1

Distributions of perceived comprehension judgements, by text, in Study 1



Appendix I: Stimulus texts used in Study 2

1) Porphyria (https://patient.info/allergies-blood-immune/porphyria-leaflet)

The porphyrias are a group of metabolic disorders. A metabolic disorder is the term used when there is a problem with one of the chemical processes within the body. In the porphyrias, the chemical process that is affected is the one that produces a substance called haem.

Haem is mainly made in the liver and in bone marrow. Haem is used to make haemoglobin which transports oxygen around the body in red blood cells. Haem is also used to make a number of proteins in the body, needed for various important functions.

There is a complex process that goes on in the liver and in bone marrow to make haem. The process has various steps and each step is controlled by a special protein called an enzyme. At each step, substances are made that are known as haem precursors. These are substances that are made during the process leading up to the making of haem. They include substances called porphyrins.

There are seven different types of porphyria. In each type, there is a lack of (or partial lack of) one of the enzymes that controls one of the steps in the making of haem. Because this enzyme is lacking, there is overproduction of haem precursors including porphyrins. The porphyrins and other precursors may then build up in the body and cause the various problems associated with porphyria. When porphyrins build up in the skin, it becomes very sensitive to sunlight and this causes the skin symptoms of porphyria. Build-up of other haem precursors in the liver and elsewhere in the body causes the symptoms that occur in the acute attacks of porphyria.

2) Scleroderma (https://www.nhs.uk/conditions/scleroderma/)

Scleroderma is an uncommon condition that results in hard, thickened areas of skin and sometimes problems with internal organs and blood vessels.

Scleroderma is caused by the immune system attacking the connective tissue under the skin and around internal organs and blood vessels. This causes scarring and thickening of the tissue in these areas.

There are several different types of scleroderma that can vary in severity. Some types are relatively mild and may eventually improve on their own, while others can lead to severe and life-threatening problems.

There's no cure for scleroderma, but most people with the condition can lead a full, productive life. The symptoms of scleroderma can usually be controlled by a range of different treatments.

Types of scleroderma and typical symptoms

There are 2 main types of scleroderma:

- localised scleroderma – just affects the skin
- systemic sclerosis – may affect blood circulation and internal organs as well as the skin

Causes of scleroderma

Normally, the body's immune system fights off any germs that infect the body. It responds like this to anything in the body it doesn't recognise, and settles down when the infection has been cleared.

It's thought scleroderma occurs because part of the immune system has become overactive and out of control. This leads to cells in the connective tissue producing too much collagen, causing scarring and thickening (fibrosis) of the tissue.

It's not clear why this happens. Certain genes are thought to be involved, and having a close family member with the condition may increase your risk.

3) Neutropenia (https://bestpractice.bmj.com/topics/en-gb/893)

Neutrophils are essential components of the haematopoietic and immune system, and quantitative or qualitative abnormalities of neutrophils can result in life-threatening infection. Neutropenia is a low neutrophil count and results from decreased production, accelerated utilisation, increased destruction, or a shift in compartments. A combination of these mechanisms may be present. Causes can be congenital or acquired. The most serious complication of neutropenia is infection, which can be fatal. The source is usually endogenous flora of the gut and mucosa (commonly Staphylococcus and gram-negative organisms). Fungal infections occur with increased frequency, but there is no increased risk of viral or parasitic infection. Common sites of infection include mucous membranes (gingivitis, stomatitis, perirectal abscesses), skin (cellulitis), and lungs (pneumonia).

Common causes

Infections are the most common causes of neutropenia in adults, followed by drug-induced neutropenias. In Europe the incidence of drug-induced neutropenia in the general population is 1.6 to 9.2 cases per million. The incidence within the US is slightly higher, with 2.4 to 15.4 cases per million reported per year. Acquired bone marrow diseases such as the leukaemias, lymphomas, and aplastic anaemia are relatively common causes of neutropenia in adults, as are nutritional deficiencies (vitamin B12, folate, copper). Other causes of neutropenia are rare in adults. The epidemiology of pseudoneutropenia (neutrophil maldistribution) is unknown.

4) Seborrhoeic dermatitis (https://www.bupa.co.uk/health-information/healthy-skin/seborrhoeic-dermatitis)

Seborrhoeic dermatitis (or seborrhoeic eczema) is a condition that makes your skin red, flaky and itchy. It affects areas of your skin that tend to be greasier, such as your face, scalp and centre of your chest.

About seborrhoeic dermatitis

Seborrhoeic dermatitis is thought to affect up to five in every 100 people in the UK. However, there are many more people who have a mild form of the disease, without even being aware they have it. Dandruff is a mild form of seborrhoeic dermatitis that affects your scalp.

You can develop seborrhoeic dermatitis at any age, from puberty onwards. It's most common to develop it between the ages of 30 and 70, although many people also get it as a teenager. It’s more common in men than women.

You can also develop a short-lived form of the condition as a baby. It mainly affects a baby's scalp, which is known as cradle cap. This usually goes away on its own within a few weeks or months.

The inflammation in seborrhoeic dermatitis is thought to be an overreaction to a yeast called Malassezia, which is present on your skin.

The symptoms vary in severity from person to person, from mild dandruff to widespread patches of red, itchy, inflamed skin. You may notice your symptoms flare up at times and then get better again.

5) Erysipelas and cellulitis (https://www.ncbi.nlm.nih.gov/books/NBK303996/)

Erysipelas and cellulitis are skin infections that can develop if bacteria enter the skin through cuts or sores. Both infections make your skin swell, become red and tender. Erysipelas (also known as St. Anthony's fire) usually only affects the uppermost layers of skin, while cellulitis typically reaches deeper layers of tissue. Provided the right treatment is started early enough, these infections usually clear up without any lasting effects. Left untreated, they sometimes lead to serious complications.

Symptoms

There are two main types of bacterial skin infections:

- Erysipelas
- Cellulitis (deeper infection of the connective tissue)

Erysipelas affects the upper layers of the skin. The typical symptom is a painful and shiny light-red swelling of a quite clearly defined area of skin. Red streaks leading from that area may be a sign that the infection has started to spread along the lymph vessels too. In more severe cases, blisters may form as well. Nearby lymph nodes sometimes swell up and become more sensitive to pressure. People usually have a fever and generally feel unwell right from the start of the infection, when the skin first turns red.

In cellulitis, the reddened skin is less clearly defined than it is in erysipelas, and it is often dark-red or slightly purplish. Unlike erysipelas, the infection caused by cellulitis reaches the lower layers of skin and the tissue beneath it. The infection can spread along tendons and muscles, and pus may form.

6) Cholestasis of pregnancy (https://www.mayoclinic.org/diseases-conditions/cholestasis-of-pregnancy/symptoms-causes/syc-20363257)

Intrahepatic cholestasis of pregnancy, commonly known as cholestasis of pregnancy, is a liver condition that occurs in late pregnancy. The condition triggers intense itching, but without a rash. Itching usually occurs on the hands and feet but can also affect other parts of the body.

Cholestasis of pregnancy can make you extremely uncomfortable. But, more worrisome are the potential complications for you and your baby. Because of the risk of complications, your doctor may recommend early delivery.

Symptoms

Intense itching is the main symptom of cholestasis of pregnancy. There is no rash. Most women feel itchy on the palms of their hands or the soles of their feet, but some women feel itchy everywhere. The itching is often worse at night and may be so bothersome that you can't sleep.

The itching is most common during the third trimester of pregnancy but sometimes begins earlier. It may feel worse as your due date approaches. Once your baby arrives, however, the itchiness usually goes away within a few days.

Causes

The cause of cholestasis of pregnancy is unclear. Your genes may play a role. Sometimes, the condition runs in families. Certain genetic variants have also been identified.

Pregnancy hormones also may be involved. Pregnancy hormones rise the closer you get to your due date. Doctors think this may slow the normal flow of bile — the digestive fluid made in the liver that helps your digestive system break down fats. Instead of leaving the liver, bile builds up in the organ. As a result, bile salts eventually enter the bloodstream, which can make you feel itchy.

7) Dengue fever (https://www.healthline.com/health/dengue-fever)

Dengue fever is a disease spread by the Aedes aegypti mosquito and is caused by one of four dengue viruses. Once you are infected with one of the dengue viruses, you will develop immunity to that virus for the rest of your life. However, you can still be infected with the other three viruses. It is possible to get all four dengue viruses in your lifetime. The viruses that cause dengue fever are related to those that cause yellow fever and West Nile virus infection.

The Centre for Disease Control and Prevention estimates that at least 400 million cases of dengue fever occur across the globe every year. Tropical regions are heavily affected.

If you contract dengue fever, symptoms usually begin about four to seven days after the initial infection. In many cases, symptoms will be mild. They may be mistaken for symptoms of the flu or another infection. Young children and people who have never experienced infection may have a milder illness than older children and adults.

Diagnosing Dengue Fever

Doctors use blood tests to check for viral antibodies or the presence of infection. If you experience dengue symptoms after traveling outside the country, you should see a healthcare provider to check if you are infected.

Treating Dengue Fever

There is no medication or treatment specifically for dengue infection. If you believe you may be infected with dengue, you should use over-the-counter pain relievers to reduce your fever, headache, and joint pain. However, aspirin and ibuprofen can cause more bleeding and should be avoided.

How to Prevent Dengue Fever

There is no vaccine to prevent dengue fever. The best method of protection is to avoid mosquito bites and to reduce the mosquito population.

8) Bilateral renal agenesis (https://www.gov.uk/government/publications/bilateral-renal-agenesis-description-in-brief/bilateral-renal-agenesis-bra-information-for-parents)

Bilateral renal agenesis is a rare condition where both kidneys do not develop.

Most babies are born with 2 kidneys, one on the left side and one on the right side. Kidneys are part of the urinary system. They filter waste and additional fluid from the blood and this is removed from the body in the form of urine (wee). Kidneys also produce hormones that help strengthen bones, control blood pressure and direct the production of red blood cells.

Kidneys form in the very early stages of a baby’s development. After around 12 weeks of pregnancy, they start to produce urine which contributes to the amniotic fluid (the water around the baby inside the womb). Babies need amniotic fluid for their lungs to develop.

As babies with bilateral renal agenesis do not have any kidneys, there is very little or no amniotic fluid at the time of the 20-week scan. No fluid or small amounts of fluid around the baby means that their lungs will not develop properly. Sadly, this means that these babies will not survive after birth. There is no way to stop or cure this condition.

Causes

We do not know exactly what causes bilateral renal agenesis, but we do know it is more common in boys. It is not caused by something you have or have not done. It is sometimes linked to other medical conditions, like those affecting your baby’s chromosomes (genetic information).

9) Zollinger-Ellison syndrome (https://rarediseases.org/rare-diseases/zollinger-ellison-syndrome/)

Zollinger-Ellison syndrome (ZES) is characterized by the development of a tumour (gastrinoma) or tumours that secrete excessive levels of gastrin, a hormone that stimulates production of acid by the stomach. Many affected individuals develop multiple gastrinomas, which are thought to have the potential to be cancerous (malignant). In most patients, the tumours arise within the pancreas and/or the upper region of the small intestine (duodenum). Due to excessive acid production (gastric acid hypersecretion), individuals with ZES may develop peptic ulcers of the stomach, the duodenum, and/or other regions of the digestive tract. Peptic ulcers are sores or raw areas within the digestive tract where the lining has been eroded by stomach acid and digestive juices. Symptoms and findings associated with ZES may include mild to severe abdominal pain; diarrhoea; increased amounts of fat in the stools (steatorrhea); and/or other abnormalities. In most affected individuals, ZES appears to develop randomly (sporadically) for unknown reasons. In approximately 25 percent of patients, ZES occurs in association with a genetic syndrome known as multiple endocrine neoplasia type 1 (MEN-1). All of the tumours are considered to have malignant potential. Prognosis is related to tumour size and the presence of distant metastases.

Affected Populations

ZES may become apparent at any age. However, symptom onset usually occurs between ages 30 and 60 years. The exact frequency of ZES in the general population is unknown. However, some researchers estimate that ZES represents less than one percent of peptic ulcers.

10) Myasthenia gravis (https://www.britannica.com/science/myasthenia-gravis)

Myasthenia gravis, chronic autoimmune disorder characterized by muscle weakness and chronic fatigue that is caused by a defect in the transmission of nerve impulses from nerve endings to muscles.

Myasthenia gravis can occur at any age, but it most commonly affects women under the age of 40 and men over the age of 60. Persons with the disease often have a higher incidence of other autoimmune disorders. Approximately 75 percent of individuals with myasthenia gravis have an abnormal thymus.

Myasthenia gravis primarily affects the muscles of the face, neck, throat, and limbs. The onset of symptoms is usually gradual, with initial manifestations of the disease seen in the muscles governing eye movements and facial expressions. Weakness may remain confined to these areas, or it may extend to other muscles, such as those involved in respiration. Muscular exertion seems to exacerbate symptoms, but rest helps restore strength.

The autoimmune reaction underlying myasthenia gravis results from a malfunction in the immune system in which the body produces autoantibodies that attack specific receptors located on the surface of muscle cells. These receptors are found at the neuromuscular junction, where nerve cells interact with muscle cells. Under normal circumstances, a nerve cell, stimulated by a nerve impulse, releases the neurotransmitter acetylcholine, which crosses the neuromuscular junction and binds to receptors on the muscle cell, thus triggering a muscular contraction. In myasthenia gravis, autoantibodies bind to the receptors, preventing acetylcholine from binding to them and thus preventing the muscle from responding to the nerve signal.

11) Coxiella burnetii infection (https://bestpractice.bmj.com/topics/en-gb/1139)

Notifiable condition in the US and some other countries.

People whose occupations put them at high risk of infection include abattoir workers, meat handlers, farmers, veterinarians, laboratory personnel, and military personnel.

Symptoms and complications correspond to either an acute infection or persistent focalised infections.

Infection during pregnancy may be associated with severe obstetric and fetal complications and endocarditis in the mother.

Acute infection can be treated with a short course of doxycycline, but persistent focalised infections require long-term therapy with doxycycline plus hydroxychloroquine. Surgical resection of infected vascular tissue or prosthetic material may also be required.

Definition

A zoonotic disease caused by the gram-negative, obligate, intracellular bacterium Coxiella burnetii. Many species of mammals, birds, and ticks are reservoirs of the bacterium, and the disease is spread globally through close contact with wild or domestic animals, especially their products of parturition, and also their urine, faeces, or milk. However, C burnetii small cell variant (pseudospores) can spread by air up to 10 kilometres from the source of infection so that exposure history is frequently lacking. Symptoms and complications are different between acute infection (i.e., a self-limiting febrile illness with varying degrees of pneumonia and hepatitis) and persistent focalised infections (e.g., endocarditis, vascular infection, osteoarticular infection, lymphadenitis). Infection during pregnancy has a specific clinical presentation (mostly asymptomatic), and may result in obstetric and fetal complications. Commonly known as Q fever.

12) Sialadenitis (https://rarediseases.org/rare-diseases/sialadenitis/)

Sialadenitis is a condition characterized by inflammation and enlargement of one or more of the salivary glands, the glands that secrete saliva into the mouth. There are both acute and chronic forms. Sialadenitis is often associated with pain, tenderness, redness, and gradual, localized swelling of the affected area. The exact cause of sialadenitis is not known.

Signs & Symptoms

Symptoms of sialadenitis include enlargement, tenderness, and redness of one or more salivary glands. These are the glands in the mouth, located near the ear (parotid), under the tongue (sublingual), and under the jaw bone (submaxillary), plus numerous small glands in the tongue, lips, cheeks and palate. Salivary stones (calculi) may block secretions from any of these glands. The gland may sometimes become infected, leading to fever and other complications.

Decreased salivary flow is a hallmark of both the acute and chronic forms of sialadenitis. The pain is more obvious while eating, and more than three-quarters of patients complain of dry mouth (xerostomia).

Causes

The exact cause of sialadenitis is unknown. In some cases, the condition may be associated with the formation of salivary gland stones (sialolithiasis).

Affected Populations

Sialadenitis affects males and females in equal numbers. It shows no racial biases.

Treatment

Initial treatment of sialadenitis involves antibiotic therapy and rehydration of the patient. Patients are referred to specialists (otolaryngologists) if any signs of facial nerve involvement are present or if drainage of the swelling is contemplated. If a stone is present, gentle massage may help move it out of the gland. Otherwise, surgery may be indicated.

13) Brucellosis (https://www.britannica.com/science/brucellosis)

Brucellosis, also called Malta fever, Mediterranean fever, or undulant fever, infectious disease of humans and domestic animals characterized by an insidious onset of fever, chills, sweats, weakness, pains, and aches, all of which resolve within three to six months. The disease is named after the British army physician David Bruce, who in 1887 first isolated and identified the causative bacteria, Brucella, from the spleen of a soldier who had died from the infection.

Humans are not natural hosts for the Brucella bacteria, and they often react violently when infected. Humans contract brucellosis either directly or indirectly from infected animals. For reasons not clearly understood, children are more resistant than adults to brucellosis. The disease is very rarely transmitted from one human being to another. In humans, acute brucellosis lasts for two weeks and then may abate, but the symptoms often return with waves of fever (from which the name undulant fever was derived) in recurring bouts of illness for about six months or a year. The infection then ceases in most people, although it can persist, sometimes for years. Chronic brucellosis is perhaps the most difficult form of the disease to diagnose because the patient's symptoms are vague and may easily be mistaken as psychological in origin. Brucellosis may be complicated by infection of the joints or spine or involvement of the heart, eyes, kidneys, or lungs. Brucella spondylitis is an arthritis of the spine that generally occurs several weeks after initial infection with brucellae and may involve any part of the spine, although the lumbar region is the most commonly affected site. The disease destroys both intervertebral disks and adjacent vertebrae but can be arrested with antibiotics and immobilization of the joints.

Appendix J: Centrality identification procedure

1) Read the Text

Fully read through the text once for the purpose of understanding.

2) Divide the Text into Coarse Idea Units

Idea units are portions of text which express a single idea. Headings may be reasonably excluded if they do not constitute an idea unit.

3) Categorise the Idea Units

Identified idea units should be categorised according to whether they are better described as either elaborative of a group of semantically-related idea units or elaborative of a specific idea unit. If the idea unit is considered to be elaborative of a semantic category of idea units, classify the idea unit as a group 1 (G1) level unit. If the idea unit is considered to be elaborative of a single idea unit, classify the idea unit in a lower grouping. For example, if the idea unit is best described as elaborating on a single idea unit at the G1-level, classify this as a group 2 (G2) level unit. Connections between successive elaborative idea units can be added on in this way: an idea unit best described as elaborating a single idea unit at the G2-level is classified as a group 3 (G3) level unit.

Categorisation of idea units should be a recursive process. Classification of subordinate idea units (below G1-level) should be reviewed where multiple G2-level (or lower) units (distributed either across different G1-level units or the same G1-level unit) are substantively semantically associated. This may suggest that these idea units are better described as G1-level units belonging to a novel idea category. A decision to reorganise should be based on the quantity of information provided by the G2-level (or lower) units about the semantic category to which they relate.

Categories of semantically related idea units at the G1-level in the context of health texts (i.e. the available slots a relevant health-condition schematic superstructure, Kintsch & van Dijk, 1978) may include:

- What the nature of the condition is; what the condition is caused by
- How is the condition developed; what causes the circumstances in which the condition occurs
- What the symptoms or signs of the condition are
- What the prevalence of the condition is
- How the condition is treated
- Whether the condition is serious

The above categories are not organised hierarchically. Novel categories of semantic related idea units may be specified which are not included on this list.

4) Select Central Idea Units and Formulate Questions

4.1) Rank Order Categories

Tally the total number of units at all grouping levels within each category, creating a ranking of the total frequency of idea units across all categories. This indicates the relative quantity of idea units within the semantic categories of information featured in the text.

4.2) Write a Summary Sentence (Macroproposition) of the G1-Level Units for Each Category

Formulate one sentence to summarise the G1-level units in the category. It should be coherent and succinct, avoid repetition, may include generalisation or may omit some G1-level units (i.e., favour brevity over exhaustiveness).

4.3) Select Ideas from the Summary Sentence of High-Ranking Categories Which are Amenable to Question Construction

It is likely a better a test of understanding of main ideas if multiple high-ranking categories are selected where multiple questions are required. For example, if 3 questions are required, testing understanding of information from 3 categories means testing a greater spread of the main ideas. However, categories will vary in their volume of units across text rankings. For example, the 3rd highest category may include three G1-level idea units on one text, or six on another text. Therefore, it may be prudent to consider the relative spread of G1-level units across categories when selecting ideas from summary sentences.

5) Select Peripheral Idea Units and Formulate Questions

Pick idea units which have comparatively little representation in the semantic structure of the text. This can be done in two ways:

- Select idea units from semantic categories with the lowest total idea unit counts. Within these categories, choose idea units that are not well connected to other idea units (not elaborated on by other idea units), this may be G1-level or lower.
- Independent of total unit counts, select the lowest level of group available.

Appendix K: Multiple-choice comprehension questions used in Study 2

For each text, the central questions are listed first (1-3) and the peripheral questions are listed second (4-6). Response options are provided in parentheses, the correct response is italicised.

1) Porphyria

1) Porphyria leads to:

   (*too many porphyrins*, too much haem, too many enzymes, too few precursors)

2) Porphyria is a problem with:

   (the bone marrow, proteins, *haem production*, metabolism)

3) Creating haem is complex because:

   (*there are many stages involved*, many different chemicals are required, the process is responsive to demand for haem, it involves multiple parts of the body)

4) Haem is used for creating proteins which:

   (act as precursors, *are used for multiple purposes*, are used to create enzymes, are involved in creating porphyrins)

5) What might relieve the skin symptoms of porphyria?

   (*covering up outside*, applying skin cream, taking medication, using sensitive-skin friendly products)

6) Having too many precursors is caused by abnormalities in the amount of:

   (porphyrins, metabolites, haem, *enzymes*)

2) Scleroderma

1) Scleroderma is caused by an immune system which is:

   (immunocompromised, under active, *over-active*, under-developed)

2) In scleroderma, what is attacked?

   (skin, collagen, *connective tissue*, vessels)

3) What is the major difference between the types of scleroderma?

(*the location of damaged tissue*, the type of damaged tissue, the frequency of tissue damage, the cause of the tissue damage)

4) The immune system works by:

(attacking invading microbes, *reacting to unfamiliar material*, checking every cell, protecting tissues from harm)

5) What causes thickened areas?

(fibrosis, *collagen*, scarring, damage)

6) Treatment for scleroderma isn't always:

(effective, available, rapid, *required*)

3) Neutropenia

1) Neutropenia is usually developed from:

(*infection*, birth, nutritional deficiencies, drugs)

2) Serious complications are caused by:

(problems with neutrophil production, abnormally functioning neutrophils, *irregularities of neutrophils*, poor concentrations of neutrophils)

3) Neutropenia can lead to:

(problems with the immune system, nutritional deficiencies, *severe infections*, poor haematopoietic function)

4) Pseudoneutropenia is a condition in which:

(neutrophils are reduced in number, neutrophils function abnormally, neutrophils are malignant, *neutrophils are unevenly spread*)

5) Neutrophils:

(*are part of the body's defences*, help metabolise drugs, neutralise toxins, are a type of neuron)

6) Infections associated with neutropenia typically occur from:

(fungal spores and moulds, contact with infectious people, *bacteria already living in our body*, multiple external sources of microbes)

4) Seborrhoeic dermatitis

1) Seborrhoeic dermatitis is a condition which develops:

   (in males, during infancy, *throughout life*, in females)

2) The symptoms of seborrhoeic dermatitis:

   (get more severe over time, are always present, won't improve without treatment, *are worse in some people*)

3) In babies, the symptoms of seborrhoeic dermatitis:

   (*will go away over time*, will become more severe, will persist through childhood, will spread to the neck and chest)

4) Seborrhoeic dermatitis is linked to a micro-organism that is found:

   (*on the body*, in fermented products, in the digestive tract, under the skin)

5) Seborrhoeic dermatitis develops on the face, neck and chest because these areas:

   (are washed frequently, are more exposed, have thinner skin, *are more oily*)

6) Why might seborrhoeic dermatitis be more common than currently thought?

   (lack of effective diagnostic testing, *the symptoms can be minimal*, many people self-treat without seeing a doctor, the flare-ups doesn't last very long)

5) Erysipelas and cellulitis

1) What kind of conditions are erysipelas and cellulitis?

   (congenital, fungal, viral, *bacterial*)

2) Erysipelas and cellulitis differ in which aspect?

   (the seriousness of the condition, the part of the body affected, *the depth of the affected tissue*, whether other tissues are involved)

3) An area affected by erysipelas or cellulitis would be:

(weeping, *enlarged*, itchy, blistered)

4) What increases your chance of developing erysipelas and cellulitis?

   (*open wounds*, poor hygiene, weakened immune system, getting older)

5) Treatment for erysipelas and cellulitis:

   (*must be prompt to be effective*, is curative with no lasting effects, only relieves the symptoms, is different for erysipelas and cellulitis)

6) In erysipelas, red streaks around the area may be a sign that:

   (*the infection is spreading*, it is developing into cellulitis, the lymphatic system is not working, pus is forming under the area)

6) Cholestasis of pregnancy

1) Cholestasis of pregnancy can primarily cause increased:

   (*scratching*, bile production, tiredness, inflammation)

2) The symptoms of cholestasis of pregnancy mostly affect the:

   (*extremities*, joints, stomach, liver)

3) The symptoms of cholestasis of pregnancy are more common:

   (before the baby develops, when the baby is first developing, *when the baby is mostly developed*, after the baby is born)

4) Bile is a substance which:

   (influences hormone production, is mostly made up of salts, *is produced by the liver*, is non-intrahepatic)

5) The symptoms of cholestasis of pregnancy can be resolved by:

   (hormone treatments, antihistamine creams, *childbirth*, having plenty of rest)

6) What causes the symptoms of cholestasis of pregnancy?

   (bile building up in the liver, increased salt content in the blood, increased pressure in the liver, *components of bile in the circulatory system*)

7) Dengue fever

1) The symptoms of dengue fever are:

   (severe in infants, *typically minimal*, develop rapidly, wide-ranging)

2) Dengue fever may be mistaken for:

   (yellow fever, malaria, West Nile virus, *influenza*)

3) Dengue fever is caused by:

   (mosquitos, *viruses*, an immune reaction, a group of infections)

4) Treatment for dengue fever involves pain killers, but:

   (*some pain medication can make symptoms worse*, effective pain relief is only available on prescription, some people can't take the pain medication, pain relievers becomes less effective over time)

5) What would reduce cases in tropical regions?

   (*controlling mosquito numbers*, limiting travel to these areas, greater vaccination uptake, better testing and isolation)

6) Testing the blood for dengue fever antibodies is done:

   (*to check if you've had dengue fever*, to identify an active case of dengue fever, when the immune system isn't fighting dengue fever, to identify the type of dengue fever)

8) Bilateral renal agenesis

1) What do babies with bilateral renal agenesis not develop?

   (hormones, amniotic fluid, a urinary tract, *multiple organs*)

2) What do the kidneys do?

   (*produce liquid waste*, regulate blood circulation, absorb and reuse excess fluid, help break down toxins and waste from blood)

3) Why is bilateral renal agenesis fatal?

(babies can't get rid of waste from their blood, *babies' lungs don't develop properly*, babies can't produce important hormones, there isn't enough amniotic fluid)

4) Bilateral renal agenesis may be related to:

   (a lack of essential nutrients in the mother's diet, exposure to harmful chemicals, a viral infection during pregnancy, *abnormalities of the DNA*)

5) Why don't kidneys produce urine before 12 weeks?

   (the amniotic fluid is not needed yet, waste has not built up in the blood yet, *the kidneys are not developed enough yet*, the baby is too small to produce urine)

6) Problems with kidney hormone production could lead to:

   (*not enough blood cells being made*, abnormal blood cells, blood cells not functioning properly, not developing the structures that produce blood cells)

9) Zollinger-Ellison syndrome

1) Zollinger-Ellison syndrome causes particular cells to:

   (stop producing proteins, *replicate out of control*, excessively produce hormones, degenerate and necrotise)

2) The abnormal cells in Zollinger-Ellison syndrome are typically found in the:

   (stomach, *pancreas*, liver, lower intestine)

3) People with Zollinger-Ellison syndrome may develop:

   (*internal lesions*, acidosis, pancreatitis, peptides)

4) The outlook for someone with Zollinger-Ellison syndrome depends on whether the affected cells:

   (have mutated, are thinly spread out over an area, have colonised the area, *are in multiple areas in the body*)

5) Someone with Zollinger-Ellison syndrome would likely experience:

   (*discomfort*, nausea, indigestion, acid reflux)

6) Which of the following is true of Zollinger-Ellison syndrome?

   (*a person with a peptic ulcer is unlikely to have Zollinger-Ellison syndrome*, a person with Zollinger-Ellison syndrome is unlikely to develop a peptic ulcer, Zollinger-Ellison syndrome is estimated to affect only one percent of people, in very few cases a peptic ulcer can lead to Zollinger-Ellison syndrome)

10) Myasthenia gravis

1) Myasthenia gravis is caused by:

   (an under-developed immune system, an immune system deficiency, *a misidentification in the immune system*, an overproduction of immune cells)

2) Myasthenia gravis causes a:

   (failure to generate nerve impulses, *disruption to neural messaging*, degeneration of muscle motor neurons, weakening of muscle tissue)

3) Someone with myasthenia gravis may have trouble with:

   (*eating*, breathing, sleeping, understanding)

4) People with myasthenia gravis are more likely to:

   (have multiple abnormal internal organs, *develop immune system problems*, be immunocompromised, have severe immune reactions)

5) The neuromuscular junctions:

   (are the muscle motor receptors, *are where nerves and muscles meet*, control the signalling of acetylcholine, control the nerve impulses)

6) In myasthenia gravis, the problem occurs because:

   (*acetylcholine can't attach to the receptor*, there is too little acetylcholine, auto-antibodies bind to acetylcholine, acetylcholine fails to trigger an impulse)

11) Coxiella burnetii infection

1) Acute and persistent focalised forms of Coxiella burnetii differ in:

(the site of initial infection, *the problems they cause*, their incubation period, the variant of the infection)

2) Coxiella burnetii is more likely to be caught from:

   (uncooked meat, dirty water, *livestock*, soil)

3) What kind of disease is Coxiella burnetii?

   (intracellular, animal, obligate, *bacterial*)

4) A doctor treating someone for Coxiella burnetii would need to:

   (get consent for treatment, *inform the appropriate authority*, determine whether the infection is acute or persistent focalised, ask about exposure history)

5) Finding the origin of the infection can be difficult because:

   (the infection can be asymptomatic, it can spread through multiple hosts before being detected, *it doesn't always require direct contact to spread*, the diagnostic tests are not completely accurate)

6) In some persistent cases of Coxiella burnetii, what may be required?

   (*replacing internal structures*, stronger doses of doxycycline, obstetric mitigation by early delivery, surgery to remove affected extremities)

12) Sialadenitis

1) Sialadenitis is frequently associated with:

   (pain in the neck, a swollen face, difficulty eating, *reduced spit production*)

2) Sialadenitis is a condition where:

   (a saliva stone has formed, the salivary gland is infected, *the salivary gland is swollen*, the salivary gland is blocked)

3) People with sialadenitis can expect to first receive:

   (therapy, drainage, *medication*, massage)

4) Across ethnicities, sialadenitis:

(*has comparable prevalence*, has different levels of severity, shows non-trivial bias, varies in the incidence of acute and chronic forms)

5) Complications may be experienced by someone with sialadenitis when:

   (they are dehydrated, *the gland is infected*, multiple glands are affected, the condition is chronic)

6) Surgery may be required if which treatment is unsuccessful?

   (*physical manipulation*, rehydration, medication, fluid drainage)

13) Brucellosis

1) Brucellosis is a disease which:

   (*affects a range of different species*, first emerged in the late 1800s, spreads easily between people, is associated with aggression)

2) Acute and chronic brucellosis infections differ in terms of the:

   (initial cause of infection, *pattern of symptoms*, development of psychological issues, type of tissues affected)

3) Brucellosis may lead to a condition which:

   (leads to muscle weakness, causes heart disease, breaks down spongy tissues, *causes inflammation of the joints*)

4) The severity of Brucellosis infection is known to be influenced by:

   (the bacterial load, *a person's age*, a person's overall health, if it was contracted directly or indirectly)

5) The brucellosis germ was first distinguished in samples taken from:

   (a soldier's swab of bodily fluid, a doctor working in the Mediterranean, *internal organ tissues*, spinal and intervertebral fluids)

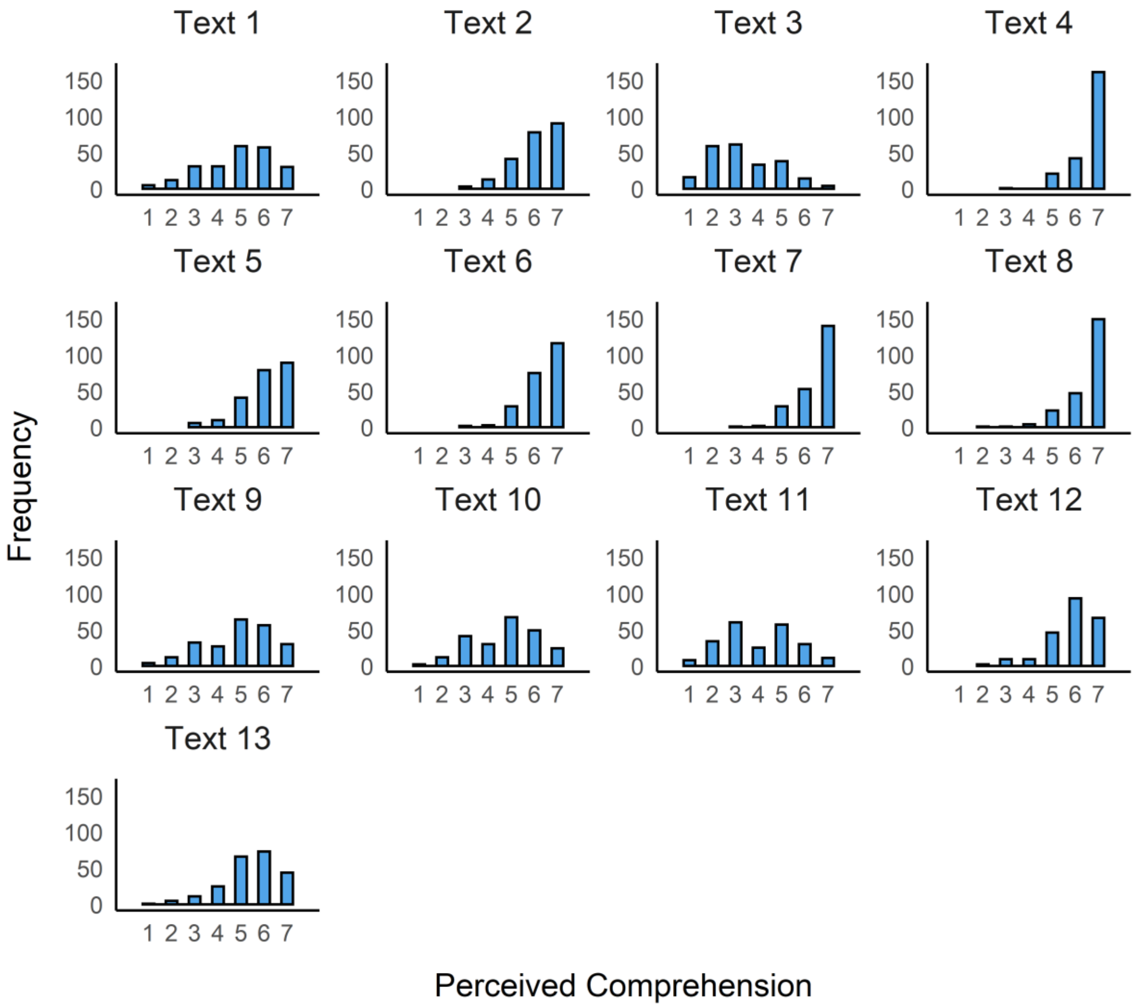
6) One of the alternative names for brucellosis captures which aspect of the infection?

(*the periodic episodes of symptoms*, the variable length of symptoms, the insidious onset of fever, the origin of the first person with the disease)

Appendix L: Supplementary figure showing judgements by text in Study 2

Figure L.1

Distributions of perceived comprehension judgements, by text, in Study 2



Appendix M: Global and local judgement simulation models used in Study 3

Global Judgement Simulation Model

Observations of the global perceived comprehension rating *G*, from individual *i* on text *j*, was the outcome of the simulation model. According to a cumulative probit model, perceived comprehension ratings are the product of a categorisation of an underlying continuous variable $\tilde{G}$, corresponding to latent perceptions of text-level comprehension. It is assumed that these underlying perceptions of comprehension are normally distributed. According to a set of *R* thresholds τ_{rij}, which partition latent perceptions of comprehension into $R + 1$ ordinal categories, global ratings are produced for each individual responding to a text. Consistent with Study 2, a 7-point ordinal response scale will be used. Correspondingly, $R = 6$ and $r = (1, \ldots, 6)$. A set of fixed population-level thresholds τ_r are assumed, with deviation in these thresholds at the level of the individual γ_i and text λ_j. Therefore, the probability P of observing a global rating G_{ij} equal to rating *r*, is given by:

$$P(G_{ij} = r) = \Phi(\tau_r) - \Phi(\tau_{r-1}),$$

where $i = 1, \ldots, N$, $j = 1, \ldots, 13$, and Φ is the cumulative distribution function of a normal distribution with a mean of 0 and a standard deviation of 1. Assuming that latent perceived comprehension $\tilde{G}$ follows a standard normal distribution, then for individual *i* reading text *j*:

$$\tilde{G}_{ij} = \eta + \varepsilon_{ij},$$

$$\eta = 0,$$

$$\varepsilon_{ij} \sim \text{Normal}(0, 1).$$

Thresholds are assumed to be a combination of fixed, population-level threshold values and variance at the level of the individual and the text. Both individual-level and text-level threshold variability are assumed to be normally distributed.

$$\tau_{rij} = \tau_r + \gamma_i + \lambda_j,$$

$$\gamma_i \sim \text{Normal}(0, \sigma_1),$$

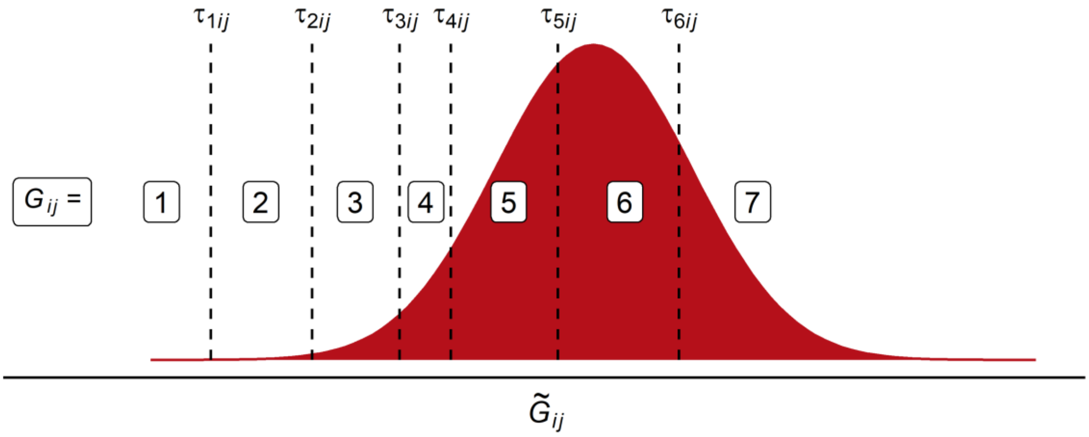
$$\lambda_j \sim \text{Normal}(0, \sigma_2).$$

Based on the model fitted to the global ratings observed in Study 2, values for τ_r (for r = 1, ..., 6) and σ_1 were selected. Rather than randomly sample observations of λ_j from the defined distribution, model-fitted estimates of the text-level deviates in thresholds were used, given that Study 3 would not require a new sample of text stimuli. The following values were selected for the thresholds τ_r, r = (1, 2, …, 6): $\tau_1 = -3.89$, $\tau_2 = -2.86$, $\tau_3 = -1.97$, $\tau_4 = -1.45$, $\tau_5 = -0.36$, $\tau_6 = 0.87$. The values for the standard deviation parameters were $\sigma_1 = 1$, and $\sigma_2 = 1.4$, however, text-level threshold deviates λ_j were not randomly sampled in the simulation. Model fitted estimates, from texts j = 1:13, were used for λ_j. These values are provided in the simulation R script in the OSF repository.

Simulated observations of global judgements of perceived comprehension were generated using the model defined above. Specifically, this was achieved by simulating observations of $\tilde{G}_{ij}$ and γ_i, then categorising $\tilde{G}_{ij}$ into ordinal ratings G_{ij} according to the thresholds τ_{rij}. This process is illustrated in Figure M.1: values of G_{ij} are obtained by categorising samples of $\tilde{G}_{ij}$ with reference to the τ_{rij} thresholds.

Figure M.1

Illustration of the Approach used to Simulate Ratings of Perceived Comprehension



Note: Positions of the τ_{rij} thresholds displayed are the expected values for *r* = (1, …, 6), given that individual γ_i and text λ_j variability in thresholds is normally distributed with mean 0.

Local Judgement Simulation Model

A local judgement of perceived comprehension L_{ik} was considered to result from the categorisation of $\tilde{L}_{ik}$, corresponding to an individual's *i* latent perception of comprehension of an idea unit *k*. Consistent with latent global perceived comprehension, $\tilde{L}$ is categorised according to a set of S thresholds τ_{sik}, to produce *S* + 1 ordinal values of *L*, where *S* = 6 and s = (1, …, 6). Therefore, for individual i reading idea k, the probability P of observing a local rating L_{ik} equal to rating s, is given by:

$$P(L_{ik} = s) = \Phi(\tau_s) - \Phi(\tau_{s-1}),$$

where i = 1, …, *N* and k = 1, …, 78. Consistent with global ratings, latent local perceived understanding was assumed to follow a standard normal distribution, such that for individual i reading idea k

$$\tilde{L}_{ik} = \eta + \varepsilon_{ik},$$

$$\eta = 0,$$

$$\varepsilon_{ik} \sim \text{Normal}(0, 1).$$

Further, the same set of population-level thresholds were employed for categorising $\tilde{L}$, as it was considered reasonable that, at the population level, individuals may evaluate the extent of their latent understanding using the same standards. With respect to variance in these thresholds, deviations from the population thresholds at the level of an individual γ_i and idea α_k were assumed to be normally distributed. Consistent with the global rating simulation model, individual-level deviates were assumed to follow a normal distribution with standard deviation σ_1.

In contrast, idea-level deviates were sampled from a normal distribution with a standard deviation σ_2 larger than that estimated for text deviates. The value of this parameter was $\sigma_2 = 2$. This choice corresponded to an assumption that idea-level variability in thresholds may be greater than text-level variability. This assumption was considered reasonable, based on the proposition that text-level deviates may capture an amalgamation of perceived understanding across multiple ideas (Dunlosky et al., 2005; Händel & Dresel, 2018; Koriat, 1995; Lefèvre and Lories, 2004), generating a smaller standard deviation of text-level deviates relative to idea-level deviates. The thresholds can, therefore, be expressed as:

$$\tau_{sik} = \tau_s + \gamma_i + \alpha_k,$$

$$\gamma_i \sim \text{Normal}(0, \sigma_1),$$

$$\alpha_k \sim \text{Normal}(0, \sigma_2).$$

Simulated observations of local judgements of perceived comprehension were generated using the model defined above. This was achieved by simulating observations of $\tilde{L}_{ik}$, γ_i and α_k, then categorising $\tilde{L}_{ik}$ into ordinal ratings L_{ik} according to the thresholds τ_{sik}. These simulated observations were sampled independently from all of those drawn in the simulation of global judgements of perceived comprehension.

Appendix N: Altered multiple-choice comprehension questions used in Study 3

Response options are provided in parentheses, the correct response is italicised. The following response options used in Study 2 were changed (where 1.3.4 refers to the fourth response option on the third question of the first text): 1.3.1, 1.3.4, 2.4.1, 3.2.1, 3.2.2, 3.2.3, 3.2.4, 7.1.3, 7.6.1, 7.6.2, 7.6.3, 8.1.2, 8.2.4, 9.1.3, 10.2.4, 10.3.1, 10.3.2, 13.1.2. In addition, the question stem for the sixth question on the seventh text was changed (7.6).

1) Porphyria

3) Creating haem is complex because:

   (*it involves many stages and parts of the body*, many different chemicals are required, the process is responsive to demand for haem, it is difficult to regulate the production of precursors)

2) Scleroderma

4) The immune system works by:

   (preventing infections in the body, *reacting to unfamiliar material*, checking every cell, protecting tissues from harm)

3) Neutropenia

2) Serious complications are caused by:

   (uncontrolled acceleration in neutrophil production, neutrophils destroying healthy tissue, *unusual behaviour or volume of neutrophils*, infection of the neutrophils)

7) Dengue fever

1) The symptoms of dengue fever are:

   (severe in infants, *typically minimal*, rapid to develop, wide-ranging)

6) A blood test for dengue fever is carried out by a doctor:

(*to examine whether viral particles are circulating*, if a person is experiencing flu-like symptoms, if a person has recently travelled to a tropical region, to identify the type of dengue fever)

8) Bilateral renal agenesis

1) What do babies with bilateral renal agenesis not develop?

   (hormones, blood cells, a urinary tract, *multiple organs*)

2) What do the kidneys do?

   (*produce liquid waste*, regulate blood circulation, absorb and reuse excess fluid, store urine)

9) Zollinger-Ellison syndrome

1) Zollinger-Ellison syndrome causes particular cells to:

   (stop producing proteins, *replicate out of control*, produce more acidic gastrin, degenerate and necrotise)

10) Myasthenia gravis

2) Myasthenia gravis causes a:

   (failure to generate nerve impulses, *disruption to neural messaging*, degeneration of muscle motor neurons, weakening of nerve tissue)

3) Someone with myasthenia gravis may have trouble with:

   (*breathing*, hearing, sleeping, understanding)

13) Brucellosis

1) Brucellosis is a disease which:

   (*affects a range of different species*, children are particularly vulnerable to, spreads easily between people, is associated with aggression)

Appendix O: Local judgement prompts used in Study 3

1) Porphyria

1) How well do you understand the impact of porphyria on the balance of porphyrins?
2) How well do you understand which substance created in the body is affected by a lack of porphyrins?
3) How well do you understand why creating haem is complex?
4) How well do you understand the function of proteins created from haem?
5) How well do you understand what could help to prevent the skin symptoms of porphyria from developing?
6) How well do you understand what causes someone to have too many precursors?

2) Scleroderma

1) How well do you understand what happens to the immune system in scleroderma?
2) How well do you understand what the immune system attacks in scleroderma?
3) How well do you understand the major difference between localised and systemic forms of scleroderma?
4) How well do you understand what a healthy immune system does?
5) How well do you understand why connective tissue cells cause thickened areas of skin?
6) How well do you understand why the treatment for scleroderma might vary?

3) Neutropenia

1) How well do you understand how people typically develop neutropenia?
2) How well do you understand the issues with neutrophils, which can cause serious complications, in someone with neutropenia?
3) How well do you understand the possible complications of neutropenia?
4) How well do you understand what happens to neutrophils in pseudoneutropenia?

5) How well do you understand the role neutrophils play in the body?
6) How well do you understand the typical source of infections associated with neutropenia?

4) Seborrhoeic dermatitis

1) How well do you understand who is at risk of developing seborrhoeic dermatitis?
2) How well do you understand differences between people in the symptoms of seborrhoeic dermatitis?
3) How well do you understand what happens to the symptoms of seborrhoeic dermatitis in babies?
4) How well do you understand where the micro-organism linked to seborrhoeic dermatitis is found?
5) How well do you understand why seborrhoeic dermatitis develops on the face, neck and chest?
6) How well do you understand why seborrhoeic dermatitis might be more common than people realize?

5) Erysipelas and cellulitis

1) How well do you understand what erysipelas and cellulitis are caused by?
2) How well do you understand the difference between erysipelas and cellulitis?
3) How well do you understand the symptoms common to both erysipelas and cellulitis?
4) How well do you understand what increases the potential of getting erysipelas or cellulitis?
5) How well do you understand what influences the success of treatments for erysipelas and cellulitis?
6) How well do you understand what it means to have red streaks on the skin in erysipelas?

6) Cholestasis of pregnancy

1) How well do you understand the main symptom of cholestasis of pregnancy?
2) How well do you understand where the main symptom of cholestasis of pregnancy is usually noticed?
3) How well do you understand when the symptoms of cholestasis of pregnancy often develop?
4) How well do you understand what bile is and what creates it?
5) How well do you understand what can resolve the symptoms of cholestasis of pregnancy?
6) How well do you understand how bile causes the main symptom of cholestasis of pregnancy?

7) Dengue fever

1) How well do you understand how most people experience the symptoms of dengue fever?
2) How well do you understand what disease has similar symptoms as dengue fever?
3) How well do you understand what causes dengue fever?
4) How well do you understand why people must be careful when selecting pain killers to treat dengue fever?
5) How well do you understand what could bring down the number of dengue fever cases in tropical regions?
6) How well do you understand why blood tests for dengue fever are carried out?

8) Bilateral renal agenesis

1) How well do you understand what does not develop in babies with bilateral renal agenesis?
2) How well do you understand what the kidneys do?

3) How well do you understand why bilateral renal agenesis is fatal?
4) How well do you understand what might be linked to developing bilateral renal agenesis?
5) How well do you understand why a baby’s kidneys do not start working earlier in pregnancy?
6) How well do you understand how blood cells can be affected by problems with kidney hormone production?

9) Zollinger-Ellison syndrome

1) How well do you understand what Zollinger-Ellison syndrome causes some cells to do?
2) How well do you understand where the abnormal cells are found in someone with Zollinger-Ellison syndrome?
3) How well do you understand what stomach acid can make a person with Zollinger-Ellison syndrome more likely to develop?
4) How well do you understand what determines the outlook for someone with Zollinger-Ellison syndrome?
5) How well do you understand the main symptoms that someone with Zollinger-Ellison syndrome might notice?
6) How well do you understand how often peptic ulcers turn out to be related to Zollinger-Ellison syndrome?

10) Myasthenia gravis

1) How well do you understand what is wrong with the immune system in myasthenia gravis?
2) How well do you understand the effect myasthenia gravis has on how the body communicates with muscles?

3) How well do you understand the symptoms that myasthenia gravis can cause?

4) How well do you understand what other conditions people with myasthenia gravis are more likely to suffer from?

5) How well do you understand what neuromuscular junctions are?

6) How well do you understand why acetylcholine does not work properly in myasthenia gravis?

11) Coxiella burnetii infection

1) How well do you understand the differences between acute and persistent focalised forms of Coxiella burnetti?

2) How well do you understand how you are most likely to catch Coviella burnetti?

3) How well do you understand what kind of infection is caused by Coxiella burnetti?

4) How well do you understand what Coxiella burnetti being a notifiable disease means?

5) How well do you understand why it can be difficult to trace the source of Coxiella burnetti infection?

6) How well do you understand the treatments that might be required for people with persistent focalised Coxiella burnetti?

12) Sialadenitis

1) How well do you understand the most distinct and typical symptom of all forms of sialadenitis?

2) How well do you understand what happens to the salivary gland in Sialadenitis?

3) How well do you understand what initial treatment will be given to people with sialadenitis?

4) How well do you understand how sialadenitis affects people of different ethnicities?

5) How well do you understand what can cause complications in sialadenitis?

6) How well do you understand why surgery may be required to treat sialadenitis?

13) Brucellosis

1) How well do you understand the characteristics of the hosts of brucella infection?
2) How well do you understand the difference between acute and chronic brucellosis infections?
3) How well do you understand what other problem can occur in the weeks after contracting brucellosis?
4) How well do you understand what influences a person's chance of severe brucellosis infection?
5) How well do you understand how brucellosis was first identified?
6) How well do you understand why brucellosis is also called 'undulant fever'?

Appendix P: Evaluation of individual-level, study-level and cross-study-level estimates of metacomprehension accuracy, given independent and identically distributed estimates.

Conditional on individual-level measures of metacomprehension accuracy being independent and identically distributed, we can expect to observe similarity in the estimates obtained in cross-study analyses of metacomprehension accuracy. To evidence this claim, the general principles supporting this claim are provided. Following this, an illustration of this is given in the context of calculating Pearson’s correlation coefficient as the individual-level estimate of metacomprehension accuracy.

General Principles

Let X_i represent a metric of metacomprehension accuracy, calculated at the level of an individual i. Assuming that observed estimates of the metric X_i are independent and identically distributed across all individuals, the distribution of X is defined for all individuals as:

$$X \sim F(\theta),$$

where F refers to a distribution function and θ refers to the parameters which define the distribution.

At the study-level, the observation of X_i from each individual is summed and then divided by the total number of participants in the sample I to produce a mean estimate of the metric of metacomprehension accuracy $\bar{\mu}_X$:

$$\bar{\mu}_X = \frac{\sum(X_i, \dots, X_I)}{I}.$$

Given the central limit theorem, the sampling distribution of the mean will be normally distributed as the number of observations approaches infinity. In the present context, as I approaches infinity:

$$\bar{\mu}_X \sim N(\mu_X, \sigma),$$

where μ_X refers to the true value of the mean of X and σ refers to the standard deviation of the sampling distribution of the mean.

At the cross-study-level, observed values of $\bar{\mu}_X$ are combined to provide an overall estimate of the value for X across participants I and studies K. Where values of $\bar{\mu}_X$ are averaged, the estimated mean of the study-level means $\bar{\bar{\mu}}_X$ is given as:

$$\bar{\bar{\mu}}_X = \frac{\sum(\bar{\mu}_{Xk}, \dots, \bar{\mu}_{XK})}{K}$$

Given the known formula for the sampling distribution of the mean for a normally distributed variable, the average of the study-level means is distributed:

$$\bar{\bar{\mu}}_X \sim N\left(\bar{\mu}_X, \frac{\sigma}{\sqrt{K}}\right).$$

Since the variance of $\bar{\bar{\mu}}_X$ is less than the variance of $\bar{\mu}_X$, conditional on I and K, cross-study-level estimates of the average value of X will have lower variability than study-level estimates of the average value of X. Further, as I and K approach infinity, the cross-study-level mean will be centred around the true mean value of X:

$$\bar{\bar{\mu}}_X \sim N\left(\mu_X, \frac{\sigma}{\sqrt{K}}\right).$$

Therefore, conditional on I and K, cross-study-level estimates are more likely to be located closer to the true average value of X than study-level estimates.

An Example: Pearson's Correlation Coefficients

Conditional on individual estimates of Pearson's correlation coefficient being independent and identically distributed, we would expect to observe greater similarity in cross-study analyses of the average coefficient than at either the individual-level or the study-level. To demonstrate this, a description is first given of the variables and approach to calculating cross-study metacomprehension accuracy. Following this, a simulation is presented to illustrate the accuracy and variability in the resulting estimates.

Defining the Variables of Interest

From each individual i, a vector of observed metacomprehension judgements $J$and a vector of comprehension scores per text S, both of length n, are collected:

$$J_i = (j_{i1}, j_{i2}, \dots, j_{in}),$$

$$S_i = (s_{i1}, s_{i2}, \dots, s_{in}).$$

For this example, assume that both judgements and comprehension scores are normally distributed, with mean μ_1 and μ_2 and standard deviation σ_1 and σ_2, respectively:

$$J \sim N(\mu_1, \sigma_1),$$

$$S \sim N(\mu_2, \sigma_2).$$

Further, assume that, jointly, J and S have a standard bivariate normal distribution with correlation coefficient ρ, such that the joint probability density function of J and S is:

$$f_{JS}(j, s) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[j^2 - 2\rho js + s^2\right]\right\}.$$

The distributions of J and S, and the correlation coefficient ρ, are assumed not to vary across individuals.

A Standard Experimental Design

From a sample of participants I within a study k, each individual participant i provides n pairs of metacomprehension judgements J and comprehension scores S. These observations are used to calculate a correlation coefficient for each individual ρ_i:

$$\rho_i = \rho(J_i, S_i).$$

Consequently, within an individual study, a vector of individual-level correlation coefficients P_k of length I is calculated:

$$P_k = (\rho_i, \dots, \rho_I).$$

To summarise the estimated correlation between metacomprehension judgements and comprehension test scores across participants I, the mean correlation coefficient $\bar{\mu}_\rho$ is calculated for study k as:

$$\bar{\mu}_{\rho k} = \frac{\sum(\rho_i, \dots, \rho_I)}{I}.$$

A Cross-Study Average

In the case where formal meta-analytical approaches to combine $\bar{\mu}_\rho$ across a set of studies K are not adopted, the mean of study-level estimated correlation coefficients may be calculated:

$$\bar{\bar{\mu}}_{\rho K} = \frac{\sum(\bar{\mu}_{\rho k}, \dots, \bar{\mu}_{\rho K})}{K}.$$

Simulated Example

In this simulated example, assume that for all participants:

$$J \sim N(3, 0.5),$$

$$S \sim N(3, 0.5),$$

$$\rho = \rho(J, S) = 0.2.$$

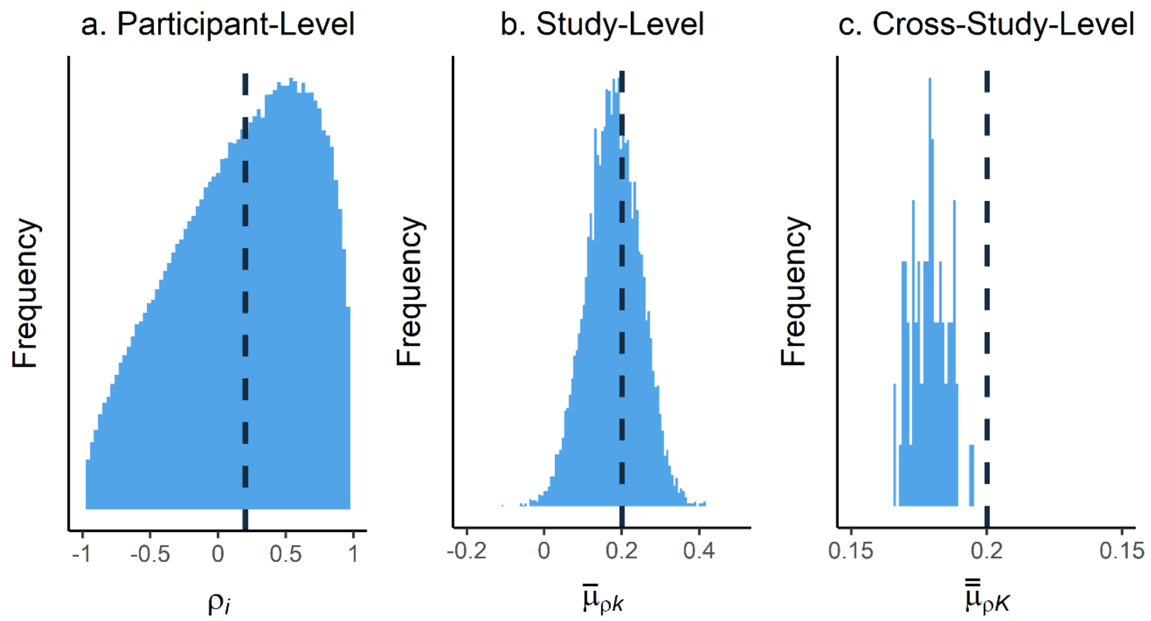
In the case where participants provide five pairs of metacomprehension judgements and comprehension scores ($n = 5$), a sample of 50 participants provides responses ($I = 50$), and the average estimated correlation coefficient from 100 studies is averaged (K = 100), values for the correlation coefficients can be simulated.

Accordingly, five observations of J and S for each participant were simulated. Pearson's correlation coefficient was calculated for each individual ρ_i, then summed across participants and divided by the total number of participants I. This produced the simulated value of the study-level estimate of the average value of the participant-level estimate $\bar{\mu}_{\rho k}$. For each set of 100 studies, the average of the study-level averages, was calculated $\bar{\bar{\mu}}_{\rho K}$. This was repeated 80 times. This can be considered analogous to a simulation of 80 meta-analyses, with each meta-analysis consisting of 100 studies, with each study sample containing 50 participants, with each participant providing five pairs of judgement-performance sores.

The distributions of the simulated values are shown in Figure P.1. Participant-level estimates of Pearson's correlation coefficient ρ_i are shown in Figure P.1a, across all of the simulated studies. Study-level estimates of the average participant-level coefficient $\bar{\mu}_{\rho k}$ are shown in Figure P.1b. Cross-study-level estimates of the mean of the study-level estimates of the average participant-level coefficient $\bar{\bar{\mu}}_{\rho K}$ are shown in Figure P.1c.

Figure P.1

Simulated Pearson's correlation coefficients at the participant-level, study-level, and cross-study-level



Note. Histograms, shaded in blue, show each estimate obtained from the simulation. The dashed line shows the true specified value of the correlation between simulated judgements and summed response accuracy (*r* = .20).

As can be seen in Figure P.1, participant-level estimates of metacomprehension are highly variable. Despite the true correlation coefficient $\rho = 0.2$, values of ρ_i span the entire range of values a correlation coefficient may take. At the study-level, however, averaged individual-level coefficients produce a distribution which is less variable and located more closely to the true correlation coefficient. Despite the reduction in variability, study-level averages nonetheless vary considerably between studies. In contrast, taking the average of study-level estimates produces mean estimates of the correlation coefficient which are typically closer to the true underlying association with considerably less variance.

Notably, as can be seen in Figure P.1b and Q.1c, averaging observations of ρ_i produces sampling distributions which are not centred on ρ. The averaged estimates of the correlation coefficients, both at the study and cross-study level are biased estimates of the true correlation coefficient. This occurs as a result of the negatively skewed distribution of individual-level coefficients correlations, as shown in Figure P.1a.

Appendix Q: Chapter 7 supplementary information

Q.1 Centrality Rating Evaluation of Idea Units

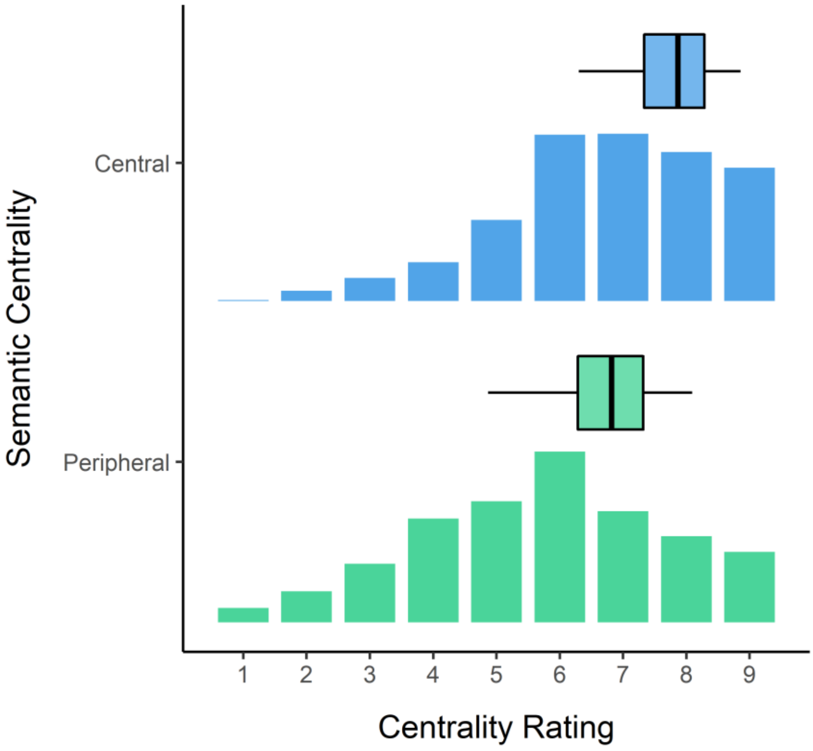
To collect centrality ratings of the 78 idea units identified from the stimulus texts, participants were recruited via Prolific. To try to ensure centrality ratings were informed by adequate comprehension of the text, the sample pool was restricted to individuals who had completed higher education. In addition, participants produced a brief written summary of each text which was evaluated to detect evidence of misunderstanding (the ratings of two participants were excluded on this basis). Participants were recruited until there were at least 20 centrality ratings for each of the 78 idea units (ratings from participants who provided partial data were also included). The sample of participants providing centrality ratings consisted of 27 individuals, 19 reporting as female and eight as male, with an average age of 37.74 (*SD* = 14.01). Participants received £2.50 for taking part.

In the centrality rating task, each text was presented on screen, followed by each of the six idea units per text, formed into coherent sentences. Participants were instructed to read the texts for the purpose of obtaining a general understanding of the condition. The rating scale appeared as an unmarked line with the labels 'Less important' on the left and 'More important' on the right of the scale. Ratings were elicited with the prompt 'Please rate the importance of the statements in understanding the text'. Previous research has utilised a variety of scale sizes to elicit centrality ratings, including four (Johnston & Afflerbach, 1982), five (Yeari et al., 2015; Yeari et al 2017), seven (Albrecht & O'Brien, 1991; Mo et al., 2007) or eight option response scales (Miller et al., 2013). As there is no clear standard, a 9-point scale was chosen to provide reasonable reliability and response discrimination (Dawes, 2008; Leung, 2011; Preston & Colman, 2000). Additional guidance was provided at the start of the experiment to clarify the rating task. These instructions stated: 'You will judge how important each statement is in understanding the health condition. You might find it helpful to read the text and try to summarise it in your head before rating. To make a rating, you could think about: if you had to explain the condition to someone else, how important it would be to include the statement in your summary; whether the text would make sense if the information in the statement was not in the text; how well connected the statement is to other parts of the text'.

The distributions of rated semantic centrality, for idea units identified as semantically central or peripheral, is shown in Figure Q.1, with boxplots displayed above each distribution to summarise the rating responses. Visually evaluating the relationship between the idea units and readers' perceptions of centrality indicated that idea units identified as central were rated as more semantically central, on average, than those identified as peripheral. However, substantial within-classification variability was observed, with the ratings of some idea units incongruous with their identification.

Figure Q.1

Ratings of Semantic Centrality for Idea Units



Note: The boxplots show the median ratings of centrality, aggregated across participants, for each of 34 idea units classified as central (blue) and each of the 34 idea units classified as peripheral (green). The histograms show the ratings of centrality aggregated across all participants and idea units, by centrality identification.

Q.2 Multiple-Choice Comprehension Questions

For each text, the central questions are listed first (1-3) and the peripheral questions are listed second (4-6). Response options are provided in parentheses, the correct response is italicised. Questions which were altered between Study 2 and 3 data collection are omitted.

1) Porphyria

1) Porphyria leads to:

   (*too many porphyrins*, too much haem, too many enzymes, too few precursors)

2) Porphyria is a problem with:

   (the bone marrow, proteins, *haem production*, metabolism)

4) Haem is used for creating proteins which:

   (act as precursors, *are used for multiple purposes*, are used to create enzymes, are involved in creating porphyrins)

5) What might relieve the skin symptoms of porphyria?

   (*covering up outside*, applying skin cream, taking medication, using sensitive-skin friendly products)

6) Having too many precursors is caused by abnormalities in the amount of:

   (porphyrins, metabolites, haem, *enzymes*)

2) Scleroderma

1) Scleroderma is caused by an immune system which is:

   (immunocompromised, under active, *over-active*, under-developed)

2) In scleroderma, what is attacked?

(skin, collagen, *connective tissue*, vessels)

3) What is the major difference between the types of scleroderma?

   (*the location of damaged tissue*, the type of damaged tissue, the frequency of tissue damage, the cause of the tissue damage)

5) What causes thickened areas?

   (fibrosis, *collagen*, scarring, damage)

6) Treatment for scleroderma isn't always:

   (effective, available, rapid, *required*)

3) Neutropenia

1) Neutropenia is usually developed from:

   (*infection*, birth, nutritional deficiencies, drugs)

3) Neutropenia can lead to:

   (problems with the immune system, nutritional deficiencies, *severe infections*, poor haematopoietic function)

4) Pseudoneutropenia is a condition in which:

   (neutrophils are reduced in number, neutrophils function abnormally, neutrophils are malignant, *neutrophils are unevenly spread*)

5) Neutrophils:

   (*are part of the body's defences*, help metabolise drugs, neutralise toxins, are a type of neuron)

6) Infections associated with neutropenia typically occur from:

   (fungal spores and moulds, contact with infectious people, *bacteria already living in our body*, multiple external sources of microbes)

4) Seborrhoeic dermatitis

1) Seborrhoeic dermatitis is a condition which develops:

(in males, during infancy, *throughout life*, in females)

2) The symptoms of seborrhoeic dermatitis:

   (get more severe over time, are always present, won't improve without treatment, *are worse in some people*)

3) In babies, the symptoms of seborrhoeic dermatitis:

   (*will go away over time*, will become more severe, will persist through childhood, will spread to the neck and chest)

4) Seborrhoeic dermatitis is linked to a micro-organism that is found:

   (*on the body*, in fermented products, in the digestive tract, under the skin)

5) Seborrhoeic dermatitis develops on the face, neck and chest because these areas:

   (are washed frequently, are more exposed, have thinner skin, *are more oily*)

6) Why might seborrhoeic dermatitis be more common than currently thought?

   (lack of effective diagnostic testing, *the symptoms can be minimal*, many people self-treat without seeing a doctor, the flare-ups doesn't last very long)

5) Erysipelas and cellulitis

1) What kind of conditions are erysipelas and cellulitis?

   (congenital, fungal, viral, *bacterial*)

2) Erysipelas and cellulitis differ in which aspect?

   (the seriousness of the condition, the part of the body affected, *the depth of the affected tissue*, whether other tissues are involved)

3) An area affected by erysipelas or cellulitis would be:

   (weeping, *enlarged*, itchy, blistered)

4) What increases your chance of developing erysipelas and cellulitis?

   (*open wounds*, poor hygiene, weakened immune system, getting older)

5) Treatment for erysipelas and cellulitis:

(*must be prompt to be effective*, is curative with no lasting effects, only relieves the symptoms, is different for erysipelas and cellulitis)

6) In erysipelas, red streaks around the area may be a sign that:

   (*the infection is spreading*, it is developing into cellulitis, the lymphatic system is not working, pus is forming under the area)

6) Cholestasis of pregnancy

1) Cholestasis of pregnancy can primarily cause increased:

   (*scratching*, bile production, tiredness, inflammation)

2) The symptoms of cholestasis of pregnancy mostly affect the:

   (*extremities*, joints, stomach, liver)

3) The symptoms of cholestasis of pregnancy are more common:

   (before the baby develops, when the baby is first developing, *when the baby is mostly developed*, after the baby is born)

4) Bile is a substance which:

   (influences hormone production, is mostly made up of salts, *is produced by the liver*, is non-intrahepatic)

5) The symptoms of cholestasis of pregnancy can be resolved by:

   (hormone treatments, antihistamine creams, *childbirth*, having plenty of rest)

6) What causes the symptoms of cholestasis of pregnancy?

   (bile building up in the liver, increased salt content in the blood, increased pressure in the liver, *components of bile in the circulatory system*)

7) Dengue fever

2) Dengue fever may be mistaken for:

   (yellow fever, malaria, West Nile virus, *influenza*)

3) Dengue fever is caused by:

(mosquitos, *viruses*, an immune reaction, a group of infections)

4) Treatment for dengue fever involves pain killers, but:

   (*some pain medication can make symptoms worse*, effective pain relief is only available on prescription, some people can't take the pain medication, pain relievers becomes less effective over time)

5) What would reduce cases in tropical regions?

   (*controlling mosquito numbers*, limiting travel to these areas, greater vaccination uptake, better testing and isolation)

8) Bilateral renal agenesis

3) Why is bilateral renal agenesis fatal?

   (babies can't get rid of waste from their blood, *babies' lungs don't develop properly*, babies can't produce important hormones, there isn't enough amniotic fluid)

4) Bilateral renal agenesis may be related to:

   (a lack of essential nutrients in the mother's diet, exposure to harmful chemicals, a viral infection during pregnancy, *abnormalities of the DNA*)

5) Why don't kidneys produce urine before 12 weeks?

   (the amniotic fluid is not needed yet, waste has not built up in the blood yet, *the kidneys are not developed enough yet*, the baby is too small to produce urine)

6) Problems with kidney hormone production could lead to:

   (*not enough blood cells being made,* abnormal blood cells, blood cells not functioning properly, not developing the structures that produce blood cells)

9) Zollinger-Ellison syndrome

2) The abnormal cells in Zollinger-Ellison syndrome are typically found in the:

   (stomach, *pancreas*, liver, lower intestine)

3) People with Zollinger-Ellison syndrome may develop:

(*internal lesions*, acidosis, pancreatitis, peptides)

4) The outlook for someone with Zollinger-Ellison syndrome depends on whether the affected cells:

   (have mutated, are thinly spread out over an area, have colonised the area, *are in multiple areas in the body*)

5) Someone with Zollinger-Ellison syndrome would likely experience:

   (*discomfort*, nausea, indigestion, acid reflux)

6) Which of the following is true of Zollinger-Ellison syndrome?

   (*a person with a peptic ulcer is unlikely to have Zollinger-Ellison syndrome*, a person with Zollinger-Ellison syndrome is unlikely to develop a peptic ulcer, Zollinger-Ellison syndrome is estimated to affect only one percent of people, in very few cases a peptic ulcer can lead to Zollinger-Ellison syndrome)

10) Myasthenia gravis

1) Myasthenia gravis is caused by:

   (an under-developed immune system, an immune system deficiency, *a misidentification in the immune system*, an overproduction of immune cells)

4) People with myasthenia gravis are more likely to:

   (have multiple abnormal internal organs, *develop immune system problems*, be immunocompromised, have severe immune reactions)

5) The neuromuscular junctions:

   (are the muscle motor receptors, *are where nerves and muscles meet*, control the signalling of acetylcholine, control the nerve impulses)

6) In myasthenia gravis, the problem occurs because:

   (*acetylcholine can't attach to the receptor*, there is too little acetylcholine, auto-antibodies bind to acetylcholine, acetylcholine fails to trigger an impulse)

11) Coxiella burnetii infection

1) Acute and persistent focalised forms of Coxiella burnetii differ in:

   (the site of initial infection, *the problems they cause*, their incubation period, the variant of the infection)

2) Coxiella burnetii is more likely to be caught from:

   (uncooked meat, dirty water, *livestock*, soil)

3) What kind of disease is Coxiella burnetii?

   (intracellular, animal, obligate, *bacterial*)

4) A doctor treating someone for Coxiella burnetii would need to:

   (get consent for treatment, *inform the appropriate authority*, determine whether the infection is acute or persistent focalised, ask about exposure history)

5) Finding the origin of the infection can be difficult because:

   (the infection can be asymptomatic, it can spread through multiple hosts before being detected, *it doesn't always require direct contact to spread*, the diagnostic tests are not completely accurate)

6) In some persistent cases of Coxiella burnetii, what may be required?

   (*replacing internal structures*, stronger doses of doxycycline, obstetric mitigation by early delivery, surgery to remove affected extremities)

12) Sialadenitis

1) Sialadenitis is frequently associated with:

   (pain in the neck, a swollen face, difficulty eating, *reduced spit production*)

2) Sialadenitis is a condition where:

   (a saliva stone has formed, the salivary gland is infected, *the salivary gland is swollen*, the salivary gland is blocked)

3) People with sialadenitis can expect to first receive:

(therapy, drainage, *medication*, massage)

4) Across ethnicities, sialadenitis:

   (*has comparable prevalence*, has different levels of severity, shows non-trivial bias, varies in the incidence of acute and chronic forms)

5) Complications may be experienced by someone with sialadenitis when:

   (they are dehydrated, *the gland is infected*, multiple glands are affected, the condition is chronic)

6) Surgery may be required if which treatment is unsuccessful?

   (*physical manipulation*, rehydration, medication, fluid drainage)

13) Brucellosis

2) Acute and chronic brucellosis infections differ in terms of the:

   (initial cause of infection, *pattern of symptoms*, development of psychological issues, type of tissues affected)

3) Brucellosis may lead to a condition which:

   (leads to muscle weakness, causes heart disease, breaks down spongy tissues, *causes inflammation of the joints*)

4) The severity of Brucellosis infection is known to be influenced by:

   (the bacterial load, *a person's age*, a person's overall health, if it was contracted directly or indirectly)

5) The brucellosis germ was first distinguished in samples taken from:

   (a soldier's swab of bodily fluid, a doctor working in the Mediterranean, *internal organ tissues*, spinal and intervertebral fluids)

6) One of the alternative names for brucellosis captures which aspect of the infection?

   (*the periodic episodes of symptoms*, the variable length of symptoms, the insidious onset of fever, the origin of the first person with the disease)

Q.3 Local Judgement Prompts

1) Porphyria

1) How well do you understand the impact of porphyria on the balance of porphyrins?

2) How well do you understand which substance created in the body is affected by a lack of porphyrins?

4) How well do you understand the function of proteins created from haem?

5) How well do you understand what could help to prevent the skin symptoms of porphyria from developing?

6) How well do you understand what causes someone to have too many precursors?

2) Scleroderma

1) How well do you understand what happens to the immune system in scleroderma?

2) How well do you understand what the immune system attacks in scleroderma?

3) How well do you understand the major difference between localised and systemic forms of scleroderma?

5) How well do you understand why connective tissue cells cause thickened areas of skin?

6) How well do you understand why the treatment for scleroderma might vary?

3) Neutropenia

1) How well do you understand how people typically develop neutropenia?

3) How well do you understand the possible complications of neutropenia?

4) How well do you understand what happens to neutrophils in pseudoneutropenia?

5) How well do you understand the role neutrophils play in the body?

6) How well do you understand the typical source of infections associated with neutropenia?

4) Seborrhoeic dermatitis

1) How well do you understand who is at risk of developing seborrhoeic dermatitis?
2) How well do you understand differences between people in the symptoms of seborrhoeic dermatitis?
3) How well do you understand what happens to the symptoms of seborrhoeic dermatitis in babies?
4) How well do you understand where the micro-organism linked to seborrhoeic dermatitis is found?
5) How well do you understand why seborrhoeic dermatitis develops on the face, neck and chest?
6) How well do you understand why seborrhoeic dermatitis might be more common than people realize?

5) Erysipelas and cellulitis

1) How well do you understand what erysipelas and cellulitis are caused by?
2) How well do you understand the difference between erysipelas and cellulitis?
3) How well do you understand the symptoms common to both erysipelas and cellulitis?
4) How well do you understand what increases the potential of getting erysipelas or cellulitis?
5) How well do you understand what influences the success of treatments for erysipelas and cellulitis?
6) How well do you understand what it means to have red streaks on the skin in erysipelas?

6) Cholestasis of pregnancy

1) How well do you understand the main symptom of cholestasis of pregnancy?
2) How well do you understand where the main symptom of cholestasis of pregnancy is usually noticed?

3) How well do you understand when the symptoms of cholestasis of pregnancy often develop?

4) How well do you understand what bile is and what creates it?

5) How well do you understand what can resolve the symptoms of cholestasis of pregnancy?

6) How well do you understand how bile causes the main symptom of cholestasis of pregnancy?

7) Dengue fever

2) How well do you understand what disease has similar symptoms as dengue fever?

3) How well do you understand what causes dengue fever?

4) How well do you understand why people must be careful when selecting pain killers to treat dengue fever?

5) How well do you understand what could bring down the number of dengue fever cases in tropical regions?

8) Bilateral renal agenesis

3) How well do you understand why bilateral renal agenesis is fatal?

4) How well do you understand what might be linked to developing bilateral renal agenesis?

5) How well do you understand why a baby's kidneys do not start working earlier in pregnancy?

6) How well do you understand how blood cells can be affected by problems with kidney hormone production?

9) Zollinger-Ellison syndrome

2) How well do you understand where the abnormal cells are found in someone with Zollinger-Ellison syndrome?

3) How well do you understand what stomach acid can make a person with Zollinger-Ellison syndrome more likely to develop?

4) How well do you understand what determines the outlook for someone with Zollinger-Ellison syndrome?

5) How well do you understand the main symptoms that someone with Zollinger-Ellison syndrome might notice?

6) How well do you understand how often peptic ulcers turn out to be related to Zollinger-Ellison syndrome?

10) Myasthenia gravis

1) How well do you understand what is wrong with the immune system in myasthenia gravis?

4) How well do you understand what other conditions people with myasthenia gravis are more likely to suffer from?

5) How well do you understand what neuromuscular junctions are?

6) How well do you understand why acetylcholine does not work properly in myasthenia gravis?

11) Coxiella burnetii infection

1) How well do you understand the differences between acute and persistent focalised forms of Coxiella burnetti?

2) How well do you understand how you are most likely to catch Coviella burnetti?

3) How well do you understand what kind of infection is caused by Coxiella burnetti?

4) How well do you understand what Coxiella burnetti being a notifiable disease means?

5) How well do you understand why it can be difficult to trace the source of Coxiella burnetti infection?

6) How well do you understand the treatments that might be required for people with persistent focalised Coxiella burnetti?

12) Sialadenitis

1) How well do you understand the most distinct and typical symptom of all forms of sialadenitis?
2) How well do you understand what happens to the salivary gland in Sialadenitis?
3) How well do you understand what initial treatment will be given to people with sialadenitis?
4) How well do you understand how sialadenitis affects people of different ethnicities?
5) How well do you understand what can cause complications in sialadenitis?
6) How well do you understand why surgery may be required to treat sialadenitis?

13) Brucellosis

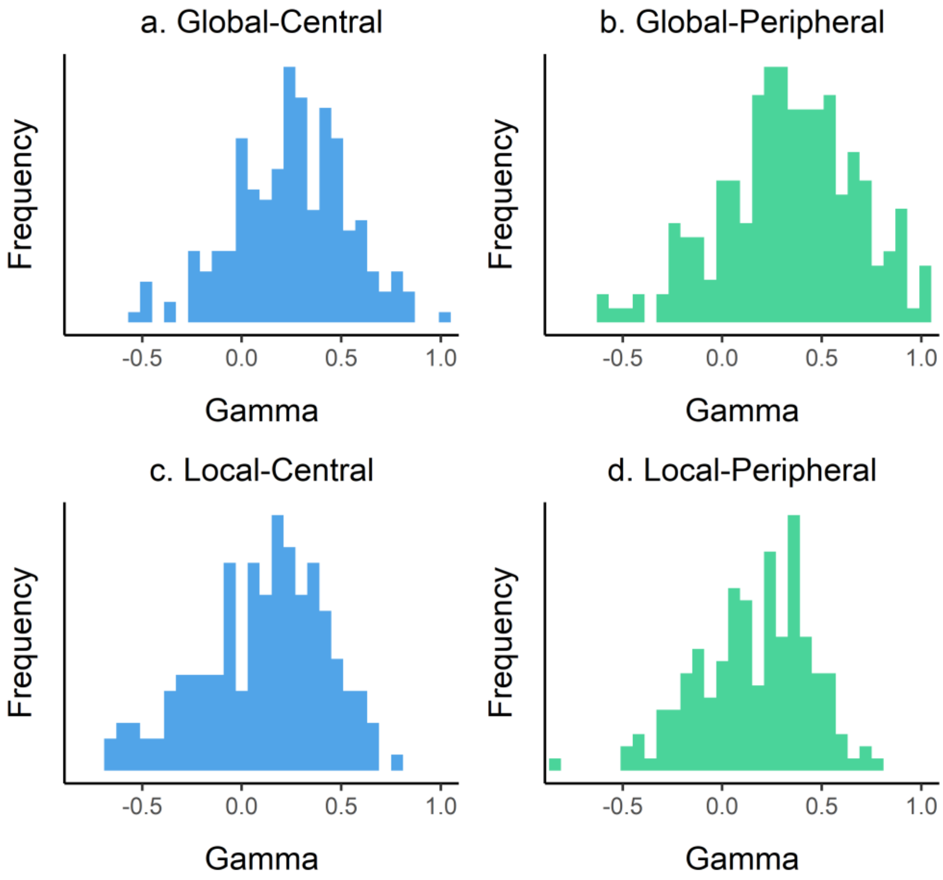
2) How well do you understand the difference between acute and chronic brucellosis infections?
3) How well do you understand what other problem can occur in the weeks after contracting brucellosis?
4) How well do you understand what influences a person's chance of severe brucellosis infection?
5) How well do you understand how brucellosis was first identified?
6) How well do you understand why brucellosis is also called 'undulant fever'?

Q.4 Individual Variation Figures

Participant-level gamma correlations, calculated in each experimental condition are shown in Figure Q.4.1. As can been seen in this figure, there is considerable variability between individuals in the estimated gamma correlations between judgements of understanding and comprehension performance.

Figure Q.4.1

Participant-Level Estimated Gamma Correlations in Each Experimental Condition

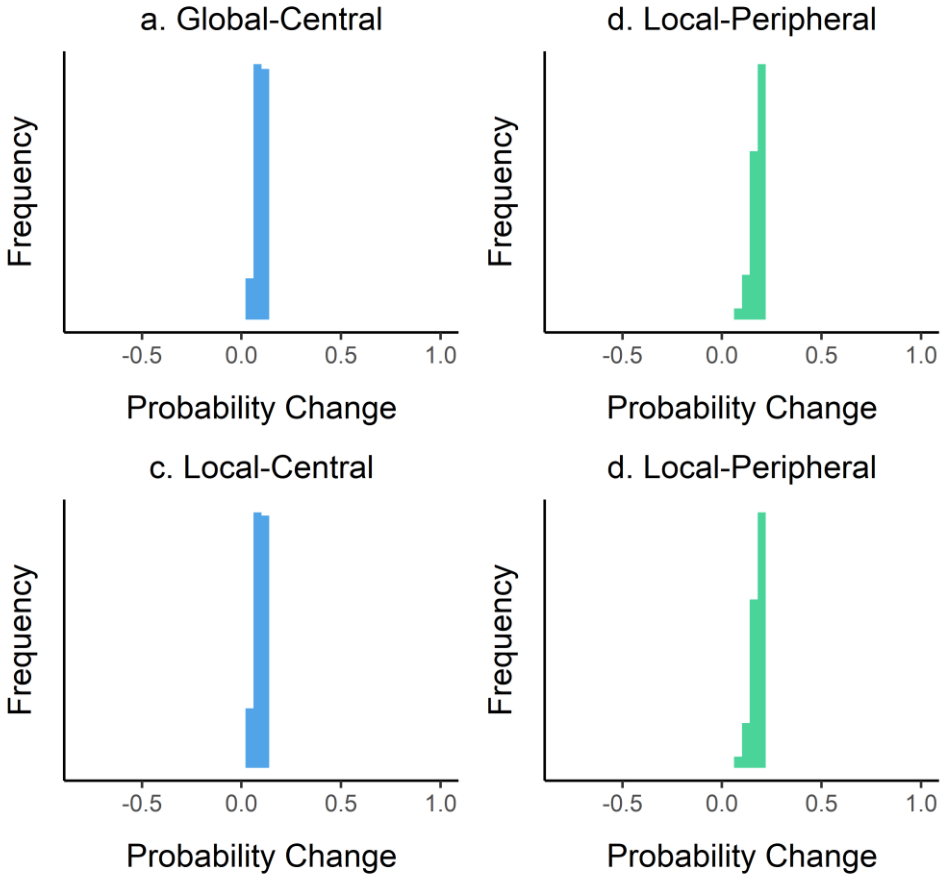


Note. Each panel shows the distribution of the estimated participant-level gamma correlations by experimental condition. Correlations are plotted by judgement scope (correlations for global judgements are shown on the top row and those for local judgements on the bottom row) and coloured by question type (central shown in blue; peripheral shown in green).

Participant-level estimated deviates, by each experimental condition, are shown in Figure Q.4.2. The deviates refer to the estimated difference, for a given participant, from the average estimated effect. As can be seen in this figure, participants are expected to show little deviation from the average relationship between judgements of understanding and comprehension performance.

Figure Q.4.2

Participant-Level Estimated Deviates in Each Experimental Condition



Note. Each panel shows the distribution of the estimated participant-level deviates in the average estimated increase in the probability of observing a correct response given a maximum judgement of understanding, compared to a minimum judgement of understanding. Deviates are plotted by judgement scope (deviates for global judgements are shown on the top row and those for local judgements on the bottom row) and coloured by question type (central shown in blue; peripheral shown in green).

Q.5 Multilevel Logistic Regression Model

The multilevel logistic regression model was fitted in R (R Core Team, 2019) using the brms package (Bürkner, 2017, 2019) and Stan (Carpenter et al., 2017), using Lancaster University's High-End Computing (HEC) facility. Weakly informative priors specified for the model parameters. For the population-level effects (the effects of perceived understanding, the scope of the judgement, the semantic centrality of the targeted information, and the interactions between these variables), normal distributions with mean 0 and standard deviation 10 were specified. Half-student-t distributions with the values of the degrees of freedom, location and scale parameters set to 3, 0 and 10 for the group-level effects (the variance terms for participants, texts and questions). An LJK correlation distribution with a shape parameter of 1 was used as the prior on the covariance between participant-level variance in the intercept and the effect of perceived comprehension. This prior allows any correlation matrix to be equally probable a priori (Lewandowski, et al., 2009).

The model was estimated using six chains with 8000 iterations each, half of which were discarded as burn-in. Model convergence was evaluated by inspecting indices of convergence, including estimates of potential scale reduction factor, autocorrelation and effective sample size, alongside visual assessment of posterior distributions and trace plots of sample chains. No issues were indicated, suggesting that the model appeared to converge well under this specification. The results of this model are presented in Table Q.5.

Table Q.5

Bayesian Multilevel Logistic Model of Response Accuracy

Parameter	Estimate[a]	Error[b]	95% CI[c]	Eff Sample
Population-level Effects				
Intercept	1.06	0.29	[0.48, 1.64]	7401
Perceived understanding	0.13	0.04	[0.06, 0.20]	20399
Judgement type	-0.07	0.09	[-0.24, 0.10]	6723
Question type	-0.55	0.29	[-1.12, 0.02]	5805
Perceived understanding x judgement type	0.01	0.05	[-0.09, 0.11]	20039
Perceived understanding x question type	0.01	0.05	[-0.08, 0.10]	22099

Judgement type x question type	0.13	0.06	[-0.01, 0.26]	38514
Perceived understanding x judgement type x question type	0.09	0.06	[-0.04, 0.21]	21150
Group-Level Variance				
Participant (intercept)	0.72	0.03	[0.66, 0.80]	8666
Participant (perceived understanding)	0.06	0.04	[0.00, 0.15]	2554
Participant (question type)	0.06	0.05	[0.00, 0.18]	2746
Text (intercept)	0.65	0.26	[0.16, 1.22]	3074
Question (intercept)	1.15	0.12	[0.94, 1.41]	5331
Covariance of intercept and slope variance				
Intercept, perceived understanding	0.03	0.34	[-0.71, 0.73]	20365
Intercept, question type	-0.01	0.43	[-0.81, 0.83]	28582
perceived understanding, question type	-0.03	0.50	[-0.89, 0.88]	7218

Note: Population-level effect estimates are presented in logits. Rhat values for all parameters = 1.00. CI = credible interval. Eff Sample = number of effective samples, obtained using the bayestestR package (Makowski et al., 2019).

[a]Estimate refers to the mean of the marginal posterior distribution of the parameter. [b]Error refers to the standard deviation of the marginal posterior distribution of the parameter. [c]Credible intervals represent the upper and lower values within which 95% of the estimated parameter values in the posterior distribution are contained.

Q.6 Multilevel Binomial Regression Model

The multilevel binomial regression model was fitted in R (R Core Team, 2019) using the brms package (Bürkner, 2017, 2019) and Stan (Carpenter et al., 2017), using Lancaster University’s High-End Computing (HEC) facility. Weakly informative priors specified for the model parameters. For the population-level effects (the effects of perceived understanding, the semantic centrality of the targeted information, and the interactions between these variables), normal distributions with mean 0 and standard deviation 10 were specified. Half-student-t distributions with the values of the degrees of freedom, location and scale parameters set to 3, 0 and 10 for the group-level effects (the variance terms for participants and texts only). An LJK correlation distribution with a shape parameter of 1 was used as the prior on the covariance between participant-level variance in the intercept and the effect of perceived comprehension. This prior allows any correlation matrix to be equally probable a priori (Lewandowski, et al., 2009).

The model was estimated using six chains with 8000 iterations each, half of which were discarded as burn-in. Model convergence was evaluated by inspecting indices of convergence, including estimates of potential scale reduction factor, autocorrelation and effective sample size, alongside visual assessment of posterior distributions and trace plots of sample chains. No issues were indicated, suggesting that the model appeared to converge well under this specification. The results of this model are presented in Table Q.6.

Table Q.6

Bayesian Multilevel Binomial Model of Response Accuracy

Parameter	Estimate[a]	Error[b]	95% CI[c]	Eff Sample
Population-level Effects				
Intercept	0.89	0.21	[0.47, 1.29]	2491
Perceived understanding	0.05	0.04	[-0.02, 0.13]	16846
Question type	-0.47	0.04	[-0.55, -0.40]	28547
Perceived understanding x question type	0.12	0.04	[0.05, 0.19]	22120
Group-Level Variance				
Participant (intercept)	0.64	0.05	[0.55, 0.73]	6693

Participant (perceived understanding)	0.05	0.03	[0.00, 0.12]	4399
Participant (question type)	0.12	0.06	[0.01, 0.23]	4955
Text (intercept)	0.70	0.17	[0.46, 1.11]	5200
Covariance of intercept and slope variance				
Intercept, perceived understanding	0.11	0.41	[-0.76, 0.84]	19578
Intercept, question type	-0.64	0.29	[-0.97, 0.17]	9072
perceived understanding, question type	-0.18	0.47	[-0.91, 0.79]	6098

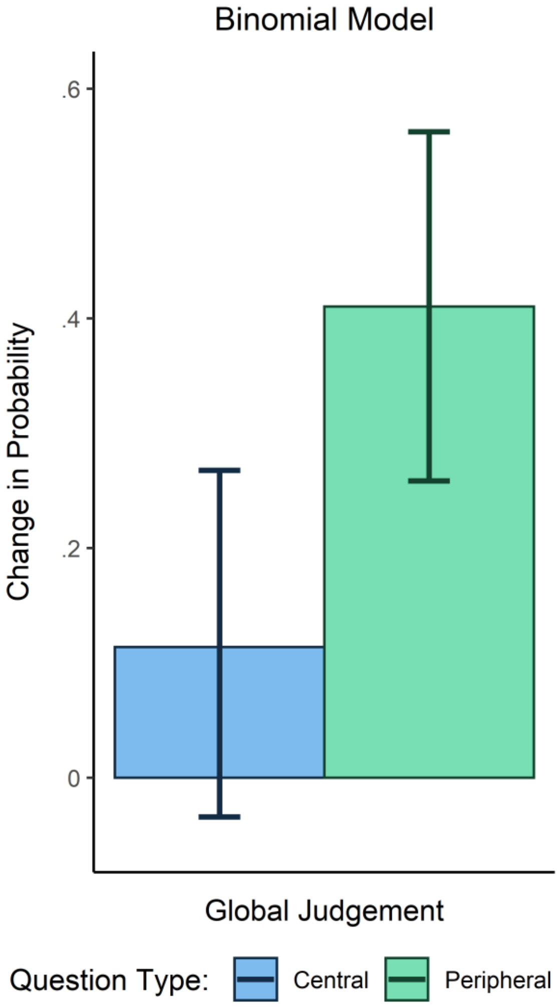
Note: Population-level effect estimates are presented in logits. Rhat values for all parameters = 1.00. CI = credible interval. Eff Sample = number of effective samples, obtained using the bayestestR package (Makowski et al., 2019).

[a]Estimate refers to the mean of the marginal posterior distribution of the parameter. [b]Error refers to the standard deviation of the marginal posterior distribution of the parameter. [c]Credible intervals represent the upper and lower values within which 95% of the estimated parameter values in the posterior distribution are contained.

The estimated effects are illustrated in Figure Q.6, showing, for each condition, the model-fitted changes in the probability of observing a correct response associated with a maximum increase in the magnitude of the judgement of understanding, with error bars displaying the 95% credible interval (i.e., the impact of a high judgement of understanding on the chance of observing evidence of understanding). The estimated beta slope was greater for semantically peripheral questions (β = 0.17) compared to semantically central questions (β = 0.05), however the 95% credible intervals for these overlapped, indicating this difference may not be robust. This can be observed in Figure Q.6, as the average estimated change differs markedly between conditions whilst the confidence intervals can be seen to overlap.

Figure Q.6

Estimated Differences Between Judgements Relating to Semantically Central and Peripheral Information in the Global Judgement Condition According to the Binomial Regression Analysis



Note. The plot shows the average estimated increase in the probability of observing a correct response given a maximum judgement of understanding, compared to a minimum judgement of understanding. Error bars represent the 95% credible interval for the mean. Estimates are coloured by question type: central shown in blue; peripheral shown in green).

9. References

Agresti, A. (2002). *Categorical Data Analysis* (2^{nd} ed.). John Wiley & Sons, Inc.

Ahmed, Y., Francis, D. J., York, M., Fletcher, J. M., Barnes, M., & Kulesz, P. (2016). Validation of the direct and inferential mediation (DIME) model of reading comprehension in grades 7 through 12. *Contemporary Educational Psychology, 44-45*, 68-82. https://doi.org/10.1016/j.cedpsych.2016.02.002

Albrecht, J. E., & O'Brien, E. J. (1991). Effects of centrality on retrieval of text-based concepts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 932-939. https://doi.org/ 10.1037/0278-7393.17.5.932

Altman, D. G., & Royston, P. (2006). The cost of dichotomising continuous variables. *BMJ, 332,* 1080-1080. https://doi.org/10.1136/bmj.332.7549.1080

Anderson, M. C. M., & Thiede, K. W. (2008). Why do delayed summaries improve metacomprehension accuracy? *Acta Psychologica*, *128*, 110-118. https://doi.org/10.1016/j.actpsy.2007.10.006

Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. *Psychological Science, 28*, 1547-1562. https://doi.org/10.1177/0956797617723724

Arbuckle, T. Y., & Cuddy, L. L. (1969). Discrimination of item strength at time of presentation. *Journal of Experimental Psychology*, *81*(1), 126-131. https://doi.org/10.1037/h0027455

Ardoin, S. P., Binder, K. S., Zawoyski, A. M., Nimocks, E., & Foster, T. E. (2019). Measuring the behaviour of reading comprehension test takers: What do they do, and should they do it? *Reading Research Quarterly, 54*(4), 507-529. https://doi.org/10.1002/rrq.246

Arnold, B. F., Hogan, D. R., Colford, J. M., & Hubbard, A. E. (2011). Simulation methods to estimate design power: An overview for applied research. *BMC Medical Research Methodology, 11,* 1-10. https://doi.org/10.1186/1471-2288-11-94

Audit Commission. (1993). *What seems to be the matter: Communication between hospital and patients*. HMSO.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390-412. https://doi.org/10.1016/j.jml.2007.12.005

Baker, J. C. M., & Dunlosky (2006). Does momentary accessibility influence metacomprehension judgments? The influence of study-judgment lags on accessibility effects. *Psychonomic Bulletin & Review, 13,* 60-65. https://doi.org/10.3758/bf03193813

Baker, L. (1979). Comprehension monitoring: Identifying and coping with text confusions. *Journal of Reading Behaviour, 11,* 365-374. https://doi.org/10.1080/10862967909547342

Baker, L. (1984). Spontaneous versus instructed use of multiple standards for evaluating comprehension: Effects of age, reading proficiency, and type of standard. *Journal of Experimental Child Psychology, 38*(2), 289–311. https://doi.org/10.1016/0022-0965(84)90127-9

Baker, L. (1985). Differences in the standards used by college students to evaluate their comprehension of expository prose. *Reading Research Quarterly, 20*(3), 297-313. https://doi.org/10.2307/748020

Baker, L. (1989). Metacognition, comprehension monitoring, and the adult reader. *Educational Psychology Review, 1,* 3-38. https://doi.org/10.1007/BF01326548

Baker. L., & Anderson, R. I. (1982). Effects of inconsistent information on text processing: Evidence for comprehension monitoring. *Reading Research Quarterly, 17*, 281-294. https://doi.org/10.2307/747487

Baker, L., & Brown, A. L. (1980). *Metacognitive skills and reading*. Center for the Study of Reading. https://files.eric.ed.gov/fulltext/ED195932.pdf

Baker, L., Zeliger-Kandasamy, A., & DeWyngaert, L. U. (2014). Neuroimaging evidence of comprehension monitoring. *Psychological Topics, 23*, 167-187.

Benjamin, A. S., & Diaz, M. (2008). Measurement of relative metamnemonic accuracy. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of metamemory and memory* (pp. 73-94). Psychology Press.

Bereiter, C., & Bird, M. (1985). Use of thinking aloud in identification and teaching of reading comprehension strategies. *Cognition and Instruction, 2*, 131–156. https://doi.org/10.1207/s1532690xci0202_2

Boehm, U., Marsman, M., Matzke, D., & Wagenmakers, E. J. (2018). On the importance of avoiding shortcuts in applying cognitive models to hierarchical data. *Behavior Research Methods, 50*, 1614-1631. https://doi.org/10.3758/s13428-018-1054-3

Bol, L., & Hacker, D. (2012). Calibration. In N. Seel (Ed.), *Encyclopedia of the sciences of learning* (pp. 495-498). https://doi.org/10.1007/978-1-4419-1428-6

Brown, A. L. (1977). *Knowing when, where, and how to remember: A problem of metacognition*. Center for the Study of Reading. https://files.eric.ed.gov/fulltext/ED146562.pdf

Brown, A. L., & DeLoache, J. S. (1978). Skills, plans, and self-regulation. In R. S. Siegler (Ed.), *Children's thinking: What develops?* (pp. 3–35). Lawrence Erlbaum Associates.

Brown, A. L., & Smiley, S. S. (1977). Rating the importance of structural units of prose passages: A problem of metacognitive development. *Child Development, 48*¸1-8. https://doi.org/10.2307/1128873

Brown, J. I., Bennett, F. M., & Hanna, G. (1981). *The Nelson Denny reading test (form E).* Riverside Publishing.

Brysbaert, M. (2019). How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of Memory and Language, 109*, 1-30. https://doi.org/10.1016/j.jml.2019.104047

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*, 1-28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P. C. (2019). Brms: Bayesian regression models using ‘Stan’ [R package]. http://CRAN.R-project.org/package=brms

Bürkner, P. C. (2022, April 11). *Estimating Non-Linear Models with brms*. The Comprehensive R Archive Network. https://cran.r-project.org/web/packages/brms/vignettes/brms_nonlinear.html

Bürkner, P. C., & Charpentier, E. (2018, November 2). Modelling monotonic effects of ordinal predictors in regression models. https://doi.org/10.31234/osf.io/9qkhj

Bürkner, P. C., & Charpentier, E. (2020). Modelling monotonic effects of ordinal predictors in Bayesian regression models. *British Journal of Mathematical and Statistical Psychology, 73*(3), 420-451. https://doi.org/10.1111/bmsp.12195

Butcher, K. R. (2006). Learning from text with diagrams: Promoting mental model development and inference generation. *Journal of Educational Psychology, 98*(1), 182–197. https://doi.org/10.1037/0022-0663.98.1.182

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munaf, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14,* 1-12. https://doi.org/10.1038/nrn3475

Calin-Jageman, R. J., & Cumming, G. (2019). The new statistics for better science: Ask how much, how uncertain, and what else is known. *The American Statistician, 73*, 271-280. https://doi.org/10.1080/00031305.2018.1518266

Calloway, R. C. (2019). *Why do you read? Toward a more comprehensive model of reading comprehension: The role of standards of coherence, reading goals, and interest* [Unpublished doctoral dissertation]. University of Pittsburgh

Cambridge University Hospitals NHS Foundation Trust (2023, February 20). *Addenbrooke's reader panel*. https://www.cuh.nhs.uk/our-services/introduction-to-patient-information/reader-panel/

Canney, G., & Winograd, P. (1979) *Schemata for reading and reading comprehension performance*. Center for the Study of Reading. https://files.eric.ed.gov/fulltext/ED169520.pdf

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betacourt, M., Brubaker, M.A., Guo, J., Li, P., & Ridell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, 76,* 1-32. https://doi.org/10.18637/jss.v076.i01

Carver, R. P. (1992). Reading rate: Theory, research, and practical implications. *Journal of Reading, 36*(2), 84-95. http://www.jstor.org/stable/40016440

Cataldo, M. G., & Oakhill, J. (2000). Why are poor comprehenders inefficient searchers? An investigation into the effects of text representation and spatial memory on the ability to locate information in text. *Journal of Educational Psychology, 92*(4), 791–799. https://doi.org/10.1037/0022-0663.92.4.791

Chadwick, S. (2018). *Individual differences in comprehension monitoring: Stability in error detection performance across age* [Unpublished master's dissertation]. Lancaster University.

Chavalarias, D., Wallach, J. D., Li, A. H. T., & Ioannidis, J. P. A. (2016). Evolution of reporting *p* values in the biomedical literature, 1990–2015. *JAMA, 315*(11), 1141-1148. https://doi.org/10.1001/jama.2016.1952

Chesterfield Royal Hospital NHS Foundation Trust. (2023, March 25). *Partnership opportunities*. https://www.chesterfieldroyal.nhs.uk/get-involved/patient-partnership/partnership-opportunities

Chiang, E., Therriault, D. J., & Franks, B. A. (2010). Individual differences in relative metacomprehension accuracy: Variation within and across task manipulations. *Metacognition and Learning, 5*(2), 121-135. https://doi.org/10.1007/s11409-009-9052-6

Christianson, K. (2016). When language comprehension goes wrong for the right reasons: Good-enough, underspecified, or shallow language processing. *The Quarterly Journal of Experimental Psychology, 69*, 817-828. https://doi.org/10.1080/17470218.2015.1134603

Christianson, K., Hollingworth, A., Halliwell, J. F., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology, 42*, 368-407. https://doi.org/ 10.1006/cogp.2001.0752

Clarke, F. R., Birdsall, T. G., & Tanner, W. P., Jr. (1959). Two types of ROC curves and definitions of parameters. Journal of the Acoustical Society of America, 31, 629-630. https://doi.org/10.1121/1.1907764

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology, 65*, 145-153. https://doi.org/10.1037/h0045186

Cohen, J. (1965). Some statistical issues in psychological research. In B.B. Wolman (Ed.), *Handbook of Clinical Psychology* (pp. 95-121). New York, NY: McGraw-Hill.

Cohen, J. (1988). *Statistical Power Analysis for Behavioural Sciences* (2nd ed.). Lawrence Erlbaum Associates.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304-1312. https://doi.org/10.1037/0003-066X.45.12.1304

Cohen, J. (1992a). A power primer. *Psychological Bulletin*, *112*, 155-159. https://doi.org/10.1037/0033-2909.112.1.155

Cohen, J. (1992b). Statistical power analysis. *Current Directions in Psychological Science, 1*, 98-101. https://doi.org/10.1111/1467-8721.ep10768783

Collins, A. A., Compton, D. L., Lindström, E. S., & Gilbert, J. K. (2020). Performance variations across reading comprehension assessments: Examining the unique contributions of text, activity, and reader. *Reading and Writing, 33*, 605-634. https://doi.org/10.1007/s11145-019-09972-5

Connor, L. T., Dunlosky, J., & Hertzog, C. (1997). Age-related differences in absolute but not relative metamemory accuracy. *Psychology and Aging, 12*, 50-71. https://doi.org/10.1037//0882-7974.12.1.50

Cromley, J. G., & Azevedo, R. (2007). Testing and refining the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology, 99*(2), 311-325. https://doi.org/10.1037/0022-0663.99.2.311

Cronbach, L. J., & Furby, L. (1970). How we should measure "change": Or should we? *Psychological Bulletin, 74*(1), 68–80. http://dx.doi.org.ezproxy.lancs.ac.uk/10.1037/h0029382

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*, 7-29. https://doi.org/10.1177/0956797613504966

Davis, T. C., Wolf, M. S., Bass, P. F., Middlebrooks, M. Kennen, E., Baker, D. W., Bennett, C. L., Durazo-Arvizu, R., Bocchini, A., Savory, S., & Parker, R. M. (2006). Low literacy impairs comprehension of prescription drug warning labels. *Journal of General Internal Medicine, 21*, 847–851. https://doi.org/10.1111/j.1525-1497.2006.00529.x

Dawes, J. (2008). Do data characteristics change according to the number of scale points used? *International Journal of Market Research, 50*, 61-77. https://doi.org/10.1177/147078530805000106

Department of Health. (2003). *Toolkit for producing patient information*. Department of Health Publications.

Downing, S. M., & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education, 10*, 61-82. https://doi.org/10.1207/s15324818ame1001_4

Dunlosky, J., Baker, J. M. C., Rawson, K. A., & Hertzog, C. (2006). Does aging influence people’s metacomprehension? Effects of processing ease on judgements of text learning. *Psychology and Aging, 21*(2), 390-400. https://doi.org/10.1037/0882-7974.21.2.390

Dunlosky, J., Hartwig, M. K., Rawson, K. A., & Lipko, A. R. (2011). Improving collefe students’ evaluations of text learning using idea-unit standards. *The Quarterly Journal of Experimental Psychology, 64*, 467-484. https://doi.org/10.1080/17470218.2010.502239

Dunlosky, J., & Lipko, A.R. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science, 16(4)*, 228-232. https://doi.org/10.1111/j.1467-8721.2007.00509.x

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. SAGE

Dunlosky, J., & Rawson, K. A. (2005). Why does rereading improve metacomprehension accuracy? Evaluating the levels-of-disruption hypothesis for the rereading effect. *Discourse Processes, 40*, 37-55. https://doi.org/ 10.1207/s15326950dp4001_2

Dunlosky, J., Rawson, K. A., & Hacker, D. J. (2002). Metacomprehension of science text: Investigating the levels-of-disruption hypothesis. In J Otero, León, J.A., & Graesser, A.C. (Eds.), The psychology of science text comprehension (pp. 255-280). Lawrence Erlbaum Associates Publishers.

Dunlosky, J., Rawson, K., & McDonald, S. (2002). Influence of practice tests on the accuracy of predicting memory performance for paired associates, sentences, and text material. In T. Perfect & B. Schwartz (Eds.), *Applied Metacognition* (pp. 68–92). Cambridge University Press.

Dunlosky, J., Rawson, K. A., & Middleton, E. L. (2005). What constrains the accuracy of metacomprehension judgements? Testing the transfer-appropriate-monitoring and accessibility hypothesis. *Journal of Memory and Language, 52*, 551-565. https://doi.org/10.1016/j.jml.2005.01.011

Eason, S. H., Goldberg, L. F., Young, K. M., Geist, M. C., & Cutting, L. E. (2012). Reader-text interactions: How differential text and question types influence cognitive skills needed for reading comprehension. *Journal of Educational Psychology, 104*, 515-528. https://doi.org/10.1037/a0027182

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of Educational Measurement* (5th ed.). Prentice-Hall.

Ehrlich, M. F., Remond, M., & Tardieu, H. (1999). Processing of anaphoric devices in young skilled and less skilled comprehenders: Differences in metacognitive monitoring. *Reading and Writing, 11*, 29-63. https://doi.org/10.1023/A:1007996502372

Embretson, S. E., & Weztel, C. D. (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement, 11*(2), 175-193. https://doi.org/10.1177/014662168701100207

Epley, N. & Gilovich, T. (2004) Are adjustments insufficient? *Personality and Social Psychology Bulletin, 30,* 447-460. https://doi.org/10.1177/0146167203261889

Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psychological Science, 17*(4), 311-318. https://doi.org/10.1111/j.1467-9280.2006.01704.x

Epstein, W., Glenberg, A. M., & Bradley, M. M. (1984). Coactivation and comprehension: Contribution of text variables to the illusion of knowing. *Memory & Cognition, 12*(4), 355-360. https://doi.org/10.3758/BF03198295

European Commission (2009). *Guideline on the readability of the labelling and package leaflet of medicinal products for human use*. https://ec.europa.eu/health/sites/default/files/files/eudralex/vol-2/c/2009_01_12_readability_guideline_final_en.pdf

Faivre, N., Vuillaume, L., Bernasconi, F., Salomon, R., Blanke, O., & Cleeremans, A. (2020). Sensorimotor conflicts alter metacognitive and action monitoring. *Cortex*, *124*, 224-234. https://doi.org/10.1016/j.cortex.2019.12.001

Fernandes, A., Malaquias, C., Figueiredo, D., da Rocha, E., & Lins, R. (2019). Why quantitative variables should not be recoded as categorical. *Journal of Applied Mathematics and Physics, 7*(7), 1519-1530. https://doi.org/10.4236/jamp.2019.77103

Ferreira, F., Christianson, K., & Hollingworth, A. (2001). Misinterpretations of garden-path sentences: Implications for models of reanalysis. *Journal of Psycholinguistic Research, 30,* 3-20. https://doi.org/10.1023/A:1005290706460

Ferreira, F., Bailey, K. G. D., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science, 11*(1), 11-15. https://doi.org/10.1111/1467-8721.00158

Ferreira, F., & Patson, N. D. (2007). The 'good-enough' approach to language comprehension. *Language and Linguistic Compass, 1*(1-2), 71-83. https://doi.org/10.1111/j.1749-818X.2007.00007.x

Ferstl, E. C., Rinck, M., von Cramon, Y. (2005). Emotional and temporal aspects of situation model processing during text comprehension: An event-related fMRI study. *Journal of Cognitive Neuroscience, 17*(5), 724-739. https://doi.org/10.1162/0898929053747658

Flavell, J. H. (1976). Metacognitive aspects of problem solving. In L. B. Resnick (Ed.), *The nature of intelligence* (pp. 231-235). Lawrence Erlbaum Associates.

Flavell, J. H. (1978). Metacognitive development. In J. M. Scandura & C. J. Brainerd (Eds.), *Structural/process theories of complex human behavior* (pp. 213-245). Sijthoff & Noordhoff

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist, 34*(10), 906-911. https://doi.org/10.1037/0003-066X.34.10.906

Flavell, J. H., Wellman, H. M. (1977). Metamemory. In R., Kail & J. Hagen, (Eds.), *Perspectives on the development of memory and cognition* (pp. 3-35). Lawrence Erlbaum Associates.

Fleming, S. (2017). HMeta-d: hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings. *Neuroscience of Consciousness, 2017*(1), 1-14. https://doi.org/10.1093/nc/nix007

Fleming, S., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience, 8*, 1-9. https://doi.org/10.3389/fnhum.2014.00443

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*, 221-233. https://doi.org/10.1037/h0057532

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019). Visualization in Bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 182*(2), 389-402. https://rss.onlinelibrary.wiley.com/doi/full/10.1111/rssa.12378

Galvin, S. J., Podd, J. V., Drga, V., & Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. *Psychonomic Bulletin & Review, 10*, 843-876. https://doi.org/10.3758/BF03196546

Garner, R. (1980). Monitoring of understanding: An investigation of good and poor readers' awareness of induced miscomprehension of text. *Journal of Reading Behaviour, 12*(1), 55-63.

Garner, R. (1987). *Metacognition and reading comprehension*. Ablex Publishing.

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science, 9*, 641-651. https://doi.org/10.1177/1745691614551642

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gernsbacher, M. A. (1990). *Language comprehension as structure building*. Lawrence Erlbaum Associates, Inc.

Gier, V. S., Kreiner, D. S., & Natz-Gonzalez, A. (2009) Harmful effects of preexisting inappropriate highlighting on reading comprehension and metacognitive accuracy, *The Journal of General Psychology, 136*(3), 287-302. https://doi.org/10.3200/GENP.136.3.287-302

Glenberg, A. M., & Epstein, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*(4), 702-718. https://doi.org/10.1037/0278-7393.11.1-4.702

Glenberg, A. M., & Epstein, W. (1987). Inexpert calibration of comprehension. *Memory & Cognition, 15*, 84-93. https://doi.org/10.3758/BF03197714

Glenberg, A. M., Sanocki, T., Epstein, W., & Morris, C. (1987). Enhancing calibration of comprehension. *Journal of Experimental Psychology: General, 116*(2), 119–136. https://doi.org/10.1037/0096-3445.116.2.119

Glenberg, A. M., Wilkinson, A. C., & Epstein, W. (1982). The illusion of knowing: Failure in the self-assessment of comprehension. *Memory & Cognition, 10*(6), 597-602. https://doi.org/10.3758/BF03202442

Golke, S., & Wittwer, J. (2017). High-performing readers underestimate their text comprehension: Artifact or psychological reality? *Proceedings of the 39th Annual Conference of the Cognitive Science Society*.

Gonzalez, R., & Nelson, T. O. (1996). Measuring ordinal association in situations that contain tied scores. *Psychological Bulletin, 119*(1), 159-165. https://doi.org/10.1037/0033-2909.119.1.159

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association, 49*, 732–764. https://doi.org/10.2307/2281536

Goodman, L. A., & Kruskal, W. H. (1963). Measures of association for cross classifications. III: Approximate sampling theory. *Journal of the American Statistical Association, 58*, 310–364. https://doi.org/10.2307/2283271

Goodman, L. A., & Kruskal, W. H. (1972). Measures of association for cross classifications. IV: Simplification of asymptotic variances. *Journal of the American Statistical Association, 67*, 415–421. https://doi.org/10.2307/2284396

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education, 7*(1), 6-10. https://doi.org/10.1177/074193258600700104

Grabe, M., Antes, J., & Kristjanson, A. (1988, Apr.). *The impact of questions and instructions on comprehension monitoring* [Paper presentation]. Meeting of the American Educational Research Association, New Orleans.

Grabe, M., Antes, J., Thorson, I. & Kahn, H. (1987). Eye fixation patterns during informed and uninformed comprehension monitoring. *Journal of Reading Behaviour, 19*(2), 123-140. https://doi.org/10.1080/10862968709547592

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers, 36*, 193-202. https://doi.org/10.3758/BF03195564

Graesser, A. C., Millis, K. K., & Zwaan, R. A. (1997). Discourse comprehension. *Annual Review of Psychology, 48*, 163–189. https://doi.org/10.1146/annurev.psych.48.1.163

Graesser, A. C., Ozuru, Y., & Sullins, J. (2009). What is a good question? In M.G. McKeown & L. Kucan (Eds.), *Bringing reading research to life* (pp. 112-141). Guilford Publications.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101*(3), 371-395. https://doi.org/10.1037/0033-295X.101.3.371

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Wiley.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J.L. Mordan (Eds.), *Syntax and semantics,* Vol. 7: *Speech acts*. Academic Press.

Griffin, T. D., Jee, B. D., & Wiley, J. (2009). The effects of domain knowledge on metacomprehension accuracy. *Memory & Cognition, 37*, 1001-1013. https://doi.org/10.3758/MC.37.7.1001

Griffin, T. D., Mielicki, M. K., & Wiley, J. (2019). Improving students' metacomprehension accuracy. In J. Dunlosky & K.A. Rawson (Eds.), *The Cambridge handbook of cognition and education* (pp. 619-646). Cambridge University Press. https://doi.org/10.1017/9781108235631

Griffin, T. D., Wiley, J., & Salas, C. R. (2013). Supporting effective self-regulated learning: The critical role of monitoring. In R. Azevedo & V. Aleven (Eds.), *International handbook of metacognition and learning technologies* (pp. 13-34). Springer.

Griffin, T. D., Wiley, J., & Thiede, K. W. (2008). Individual differences, rereading, and self-explanation: Concurrent processing and cue validity as constraints on metacomprehension accuracy. *Memory & Cognition, 36*, 93-103. https://doi.org/10.3758/mc.36.1.93

Griffin, T. D., Wiley, J., & Thiede, K. W. (2019). The effects of comprehension-test expectancies on metacomprehension accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*(6), 1066–1092. https://doi.org/10.1037/xlm0000634

Hacker, D. J. (1998). Self-regulated comprehension during normal reading. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 165-191). Lawrence Erlbaum Associates Publishers.

Händel, M., de Bruin, A. B. H., & Dresel (2020). Individual differences in local and global metacognitive judgements. *Metacognition and Learning, 15*, 51-75. https://doi.org/10.1007/s11409-020-09220-0

Händel, M., & Dresel, M. (2018). Confidence in performance judgement accuracy: The unskilled and unaware effect revisited. *Metacognition and Learning, 13*, 265-285. https://doi.org/10.1007/s11409-018-9185-6

Harrell, F. E., & Slaughter, J. C. (2020). *Biostatistics for biomedical research*. https://hbiostat.org/doc/bbr.pdf

Hart, J. T. (1967). Memory and the memory-monitoring process. *Journal of Verbal Learning & Verbal Behavior, 6*(5), 685-691. https://doi.org/10.1016/S0022-5371(67)80072-0

Hattie, J. (2013). Calibration and confidence: Where to next? *Learning and Instruction, 24*, 62-66. http://doi.org/10.1016/j.learninstruc.2012.05.009

Hedges, L. V., & Rhoads, C. (2010). Statistical power analysis. In P. Peterson, E. Baker, & B. McGaw (Eds.) *International Encyclopedia of Education* (pp. 436-443). Elsevier Ltd. https://doi.org/10.1016/B978-0-08-044894-7.01356-7

Helder, A., Perfetti, C. A., van den Broek, P., Stafura, J. Z., & Calloway, R. C. (2019). ERP indicators of local and global text influences on word-to-text integration. *Language, Cognition and Neuroscience*, *34*, 13-28. https://doi.org/10.1080/23273798.2018.1496268

Helder, A., van den Broek, P., Karlsson, J., & Van Leijenhorst, L. (2017). Neural correlates of coherence-break detection during reading of narratives. *Scientific Studies of Reading, 21*(6), 463–479. https://doi.org/10.1080/10888438.2017.1332065

Helder, A., Van Leijenhorst, L., & van den Broek, P. (2016). Coherence monitoring by good and poor comprehenders in elementary school: Comparing offline and online measures. *Learning and Individual Differences, 48*, 17–23. https://doi.org/10.1016/j.lindif.2016.02.008

Higham, P. A. (2007). No special K! A signal detection framework for the strategic regulation of memory accuracy. *Journal of Experimental Psychology: General, 136*(1), 1–22. https://doi.org/10.1037/0096-3445.136.1.1.

Higham, P. A., & Higham, D. P. (2019). New improved gamma: Enhancing the accuracy of Goodman-Kruskal's gamma using ROC curves. *Behavior Research Methods, 51*, 105-125. https://doi.org/10.3758/s13428-018-1125-5

Hilbe, J. M. (2009). *Logistic Regression Models*. CRC Press. https://doi.org/10.1201/9781420075779

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E. J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review, 21*, 1157-1164. https://doi.org/ 10.3758/s13423-013-0572-3

Huff, J. D., & Nietfeld, J. L. (2009). Using strategy instruction and confidence judgments to improve metacognitive monitoring. *Metacognition and Learning*, *4*(2), 161–176. https://doi.org/10.1007/s11409-009-9042-8

Humphreys, K., & Weisner, C. (2000). Use of exclusion criteria in selecting research subjects and its effect on the generalizability of alcohol treatment outcome studies. *American Journal of Psychiatry, 157*(4), 588-594. https://doi.org/10.1176/appi.ajp.157.4.588

Institute of Medicine (2004). *Health literacy: A prescription to end confusion*. The National Academies Press. https://doi.org/10.17226/10883

Jaeger, A. J., & Wiley, J. (2014). Do illustrations help or harm metacomprehension accuracy? *Learning and Instruction, 34*, 58-73. https://doi.org/10.1016/j.learninstruc.2014.08.002

Jee, B., Wiley, J., & Griffin, T. (2006). Expertise and the illusion of comprehension. *Proceedings of the Annual Conference of the Cognitive Science Society*.

Jiroutek, M. R., Muller, K. E., Kupper, L. L., & Steward, P. W. (2003). A new method for choosing sample size for confidence interval-based inferences. *Biometrics, 59*(3), 580-590. https://doi.org/10.1111/1541-0420.00068

Johnson, P. C. D., Barry, S. J. E., Ferguson, H. M., & Müller, P. (2015). Power analysis for generalized linear mixed models in ecology and evolution. *Methods in Ecology and Evolution, 6*, 133-142. https://doi.org/10.1111/2041-210X.12306

Johnson, R. E. (1970). Recall of prose as a function of the structural importance of the linguistic units. *Journal of Verbal Learning & Verbal Behavior, 9*(1), 12–20. https://doi.org/10.1016/S0022-5371(70)80003-2

Johnston, P., & Afflebacher, P. (1982, Nov.). *Centrality and reading comprehension test questions* [Paper presentation]. Annual Meeting of the New York State Reading Association, New York.

Kayarkaya, B., & Unaldi, A. (2020). What you might not be assessing through a multiple choice test task. *International Journal of Assessment Tools in Education, 7,* 98-113. https://dx.doi.org/10.21449/ijate.699494

Keener, M. C., & Hacker, D. J. (2012). Comprehension monitoring. In N. Seel (Ed.), *Encyclopedia of the sciences of learning* (pp. 691-693). https://doi.org/10.1007/978-1-4419-1428-6

Kelley, K., Maxwell, S. E., & Rausch, J. R. (2003). Obtaining power or obtaining precision: Delineating methods of sample-size planning. *Evaluation & The Health Professions, 26*(3), 258-287. https://doi.org/10.1177/0163278703255242

Kendall, M. G. (1955). *Rank Correlation Methods* (2nd ed.). London: Charles Griffin.

Kendeou, P., Savage, R., & van den Broek, P. (2009). Revisiting the simple view of reading. *British Journal of Educational Psychology, 79*, 353-370. https://doi.org/10.1348/978185408X369020

Kendeou, P., & van den Broek, P. (2005) The effects of readers' misconceptions on comprehension of scientific text. *Journal of Educational Psychology, 97*, 235-245. https://doi.org/10.1037/0022-0663.97.2.235

Kickbush, I. S. (2001). Health literacy: Addressing the health and education divide. *Health Promotion International, 16*(3), 289-297. https://doi.org/10.1093/heapro/16.3.289

Kim, O. Kendeou, P., van den Broek, P., White, M. J., & Kremer, K. (2008). Cat, rat, and rugrats: Narrative comprehension in young children with Down Syndrome. *Journal of Developmental Physical Disabilities, 20*, 337-351. https://doi.org/10.1007/s10882-008-9101-0

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95*, 163-182. https://doi.org/10.1037/0033-295X.95.2.163

Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist, 49*, 294-303. https://doi.org/10.1037/0003-066X.49.4.294

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.

Kintsch, W. (2012). Psychological models of reading comprehension and their implications for assessment. In J. P. Sabatini, E. R. Albro, & T. O'Reilly (Eds.), *Measuring up:*

Advances in how to assess reading ability (pp. 21-38). Rowman & Littlefield Education.

Kintsch, W., & Rawson, K. A. (2004). Comprehension. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp.211-226). Blackwell Publishing. https://doi.org/10.1002/9780470757642.ch12

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*(5), 363-394. https://doi.org/10.1037/0033-295X.85.5.363

Kools M., van de Wiel M. W., Ruiter R. A., Crüts A., & Kok G. (2006). The effect of graphic organizers on subjective and objective comprehension of a health education text. *Health Education & Behavior, 33*(6), 760-772. https://doi.org/10.1177/1090198106288950

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review, 100*, 609-639. https://doi.org/10.1037/0033-295X.100.4.609

Koriat, A. (1995). Dissociating knowing and the feeling of knowing: Further evidence for the accessibility model. *Journal of Experimental Psychology: General, 124*, 311-333. https://doi.org/10.1037/0096-3445.124.3.311

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*(4), 349-370. https://doi.org/10.1037/0096-3445.126.4.349

Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Predicting one’s own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General, 133*, 643-656. https://doi.org/10.1037/0096-3445.133.4.643

Kristensen, S. B., Sandberg, K., Bibby, B. M. (2020). Regression methods for metacognitive sensitivity. *Journal of Mathematical Psychology, 94*, 1-17. https://doi.org/10.1016/j.jmp.2019.102297

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. Journal of *Personality and Social Psychology, 77*(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Krushcke, J. K. (2013). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General, 142*(2), 573-603. https://doi.org/10.1037/a0029146

Kruschke, J. K. (2015). *Doing Bayesian Analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). Academic Press.

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review, 25*, 178-206. https://doi.org/10.3758/s13423-016-1221-4

Kubik, V., Jemstedt, A., Eshratabadi, H. M., Schwartz, B., & Jönsson, F. U. (2022). The underconfidence-with-practice effect in action memory: The contribution of retrieval practice to metacognitive monitoring. *Metacognition and Learning, 17*, 375-398. https://doi.org/10.1007/s11409-021-09288-2

Kuczera, M., Field, S., & Windisch, H. C. (2016). *Building skills for all: A review of England*. OECD. https://www.oecd.org/unitedkingdom/building-skills-for-all-review-of-england.pdf

Kulesz, P. A., Francis, D. J., Barnes, M., & Fletcher, J. M. (2016). The influence of properties and their interactions with reader characteristics on reading comprehension:

An explanatory item response study. *Journal of Educational Psychology, 108*(8), 1078-1097. https://doi.org/10.1037/edu0000126

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience, 31*(1), 32–59. https://doi.org/10.1080/23273798.2015.1102299

Kurby, C. A., Ozuru, Y., & McNamara, D. S. (2007). Individual differences in comprehension monitoring ability during reading. *Proceedings of the Annual Meeting of the Cognitive Science Society.*

Kwon, H., & Linderholm, T. (2014). Effects of Self-Perception of Reading Skill on Absolute Accuracy of Metacomprehension Judgements. *Current Psychology, 33*, 73-88. https://doi.org/10.1007/s12144-013-9198-x

Lambert, B. (2018). A *Student's Guide to Bayesian Statistics*. Sage

Landau, S., & Stahl, D. (2013). Sample size and power calculations for medical studies by simulation when closed form expressions are not available. *Statistical Methods in Medical Research, 22*(3), 324-345. https://doi.org/10.1177/0962280212439578

Lee, S. Y. D., Stucky, B. D., Lee, J. Y., Rozier, R. G., & Bender, D. E. (2010). Short Assessment of Health Literacy – Spanish and English: A comparable test of health literacy for Spanish and English speakers. *Health Services Research, 45*(4), 1105-1120. https://doi.org/10.1111/j.1475-6773.2010.01119.x

Lefèvre, N., & Lories, G. (2004). Text cohesion and metacomprehension: Immediate and delayed judgements. *Memory & Cognition, 32,* 1238-1254. https://doi.org/10.3758/bf03206315

Lehman, S., & Schraw, G. (2002). Effects of coherence and relevance on shallow and deep text processing. *Journal of Educational Psychology, 94*, 738-750. https://doi.org/10.1037//0022-0663.94.4.738

Leslie, L., & Caldwell, J. S. (2017). *Qualitative reading inventory* (6th ed.). Pearson.

Leung, S. O. (2011). A comparison of psychometric properties and normality in 4-, 5-, 6-, and 11- point Likert scales. *Journal of Social Service Research, 37*, 412-421. https://doi.org/10.1080/01488376.2011.580697

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis, 100*(9), 1989-2001. https://doi.org/10.1016/j.jmva.2009.04.008

Lin, L. M., Moore, D., & Zabrucky, K. M. (2000). Metacomprehension knowledge and comprehension of expository and narrative texts among younger and older adults. *Educational Gerontology*, *26*(8), 737-749. https://doi.org/10.1080/036012700300001395

Lin, L. M., Moore, D., & Zabrucky, K. M. (2001). An assessment of students' calibration of comprehension and calibration of performance using multiple measures. *Reading Psychology, 22*(2), 111–128. https://doi.org/10.1080/027027101300213083

Lin, L. M., & Zabrucky, K. M. (1998). Calibration of comprehension: Research and implications for education and instruction. *Contemporary Educational Psychology, 23*, 345-391. https://doi.org/10.1006/ceps.1998.0972

Lin, L. M., Zabrucky, K. M., & Moore, D. (2002). Effects of text difficulty and adults' age on relative calibration of comprehension. *The American Journal of Psychology, 115*(2), 187-198. https://doi.org/10.2307/1423434

Linderholm, T., Wang, X., Therriault, D., Zhao, Q., & Jakiel, L. (2012). The accuracy of metacomprehension judgements: The biasing effect of text order. *Electronic Journal of Research in Educational Psychology, 10*(1), 111-128. https://doi.org/10.25115/ejrep.v10i26.1487

Linderholm, T., & Wilde, A. (2010). College students' beliefs about comprehension when reading for different purposes. *Journal of College Reading and Learning, 40*, 7-19. https://doi.org/10.1080/10790195.2010.10850327

Linderholm, T., Zhao, Q., Therriault, D. J., & Cordell-McNulty, K. (2008). Metacomprehension effects situated within an anchoring and adjustment framework. *Metacognition Learning, 3,* 175-188. https://doi.org/10.1007/s11409-008-9025-1

Livingston, J. A. (2003). Metacognition: An overview. *U.S. Department of Education*. http://files.eric.ed.gov/fulltext/ED474273.pdf

Löffler, E., von der Linden N., & Schneider, W. (2016). Influence of domain knowledge on monitoring performance across the life span. *Journal of Cognition and Development, 17*(5), 765-785. https://doi.org/10.1080/15248372.2016.1208204

Lord, F. M. (1956). The measurement of growth. *Educational and Psychological Measurement, 16*, 321-437. http://dx.doi.org/10.1177/001316445601600401.

Lynch, J. S., van den Broek, P., Kremer, K. E., Kendeou, P., White, M. J., & Lorch, E. P. (2008). The development of narrative comprehension and its relation to other early reading skills. *Reading Psychology, 29*, 327–365. https://doi.org/ 10.1080/02702710802165416

MacGinitie, W. H., & MacGinitie, R. K. (1989). *Gates-MacGinitie reading tests* (3rd ed). The Riverside Publishing Company.

Madison, E. M., & Fulton, E. K. (2022). The influence of summary modality on metacomprehension accuracy. *Metacognition and Learning, 17*, 117-138. https://doi.org/10.1007/s11409-021-09277-5

Magliano, J. P., Little, L. D., & Graesser, A. C. (1993). The impact of comprehension instruction on the calibration of comprehension. *Reading Research and Instruction, 32*(3), 49–63. https://doi.org/10.1080/19388079309558124

Maki, R. H. (1998). Metacomprehension of text: Influence of absolute confidence level on bias and accuracy. In D. L. Medin (Ed.), *The psychology of learning and motivation: Advances in research and theory*, Vol. 38, (pp. 223–248). Academic Press.

Maki, R. H., & Berry, S. L. (1984). Metacomprehension of text material. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(4), 663–679. https://doi.org/10.1037/0278-7393.10.4.663

Maki, R. H., Foley, J. M., Kajer, W. K., Thompson, R. C., & Willert, M. G. (1990). Increased processing enhances calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(4), 609-616. https://doi.org/10.1037/0278-7393.16.4.609

Maki, R. H., & Serra, M. (1992). The basis of test predictions for text material. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(1), 116–126. https://doi.org/10.1037/0278-7393.18.1.116

Maki, R. H., Shields, M., Wheeler, A. E., Z., & Zacchilli, T. L. (2005). Individual differences in absolute and relative metacomprehension accuracy. *Journal of Educational Psychology, 97*(4), 723-731. https://doi.org/10.1037/0022-0663.97.4.723

Makowski, D., Ben-Shachar, M., & Lüdecke, D. (2019). BayestestR: Describing effects and their uncertainty, existence and significance within the bayesian framework. *Journal of Open Source Software, 4*, 1-8. https://doi.org/10.21105/joss.01541

Mandler, J. M., & Johnson, N. S. (1977). Remembrance of things parsed: Story structure and recall. *Cognitive Psychology, 9*, 111-151. https://doi.org/10.1016/0010-0285(77)90006-8

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition, 21*(1), 422–430. https://doi.org/10.1016/j.concog.2011.09.021

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition: An International Journal, 21*(1), 422–430. https://doi.org/10.1016/j.concog.2011.09.021

Maniscalco, B., & Lau, H. (2014). Signal Detection Theory analysis of type 1 and type 2 data: Meta-*d'*, response-specific meta-*d'*, and the unequal variance SDT model. In S. M. Fleming & C. D. Frith (Eds.), *The cognitive neuroscience of metacognition* (pp. 25-66). Springer Berlin. https://doi.org/10.1007/978-3-642-45190-4_3

Margolin, S. J. (2013). Can bold typeface improve readers' comprehension and metacomprehension of negation? *Reading Psychology, 34*(1), 85–99. https://doi.org/10.1080/02702711.2011.626107

Markman, E. M. (1977). Realizing that you don't understand: A preliminary investigation. *Child Development, 48,* 986-992. https://doi.org/10.2307/1128350

Markman, E. M. (1979). Realizing that you don't understand: Elementary school children's awareness of inconsistencies. *Child Development, 50,* 643-655. https://doi.org/10.2307/1128929

Masson, M. E. J., & Rotello, C. M. (2009). Sources of bias in the Goodman–Kruskal gamma coefficient measure of association: Implications for studies of metacognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(2), 509–527. https://doi.org/10.1037/a0014876

Maxwell, S.E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9,* 147-163. https://doi.org/10.1037/1082-989X.9.2.147

Maxwell, S. E., Kelley, K., & Rausch, J. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology, 59*(1), 537-563. https://doi.org/10.1146/annurev.psych.59.103006.093735

Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). Cambridge University Press

Mazancieux, A., Dinze, C., Souchay, C., & Moulin, C. J. A. (2020). Metacognitive domain specificity in feeling-of-knowing but not retrospective confidence. *Neuroscience of Consciousness, 6*(1), 1-11. https://doi.org/10.1093/nc/niaa001

McCabe, D. P., & Castel, A. D. (2008). Seeing is believing: The effect of brain images on judgments of scientific reasoning. *Cognition, 107*(1), 342-352. https://doi.org/10.1016/j.cognition.2007.07.017

McConnell, B., & Vera-Hernández, M. (2015). *Going beyond simple sample size calculations: A practitioner's guide* (Report No. W15/17). Institute for Fiscal Studies. https://doi.org/10.1920/wp.ifs.2015.1517

McCrudden, M. T., Magliano, J. P., & Schraw, G. (2011). Toward an integrated view of relevance in text comprehension. In M.T. McCrudden, J.P. Magliano, & G. Schraw (Eds.), *Text relevance and learning from text* (pp. 395–414). Information Age.

McKoon, G., & Ratcliff, R. (1992). Inference during reading. *Psychological Review, 99*(3), 440–466. https://doi.org/10.1037/0033-295X.99.3.440

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*(4), 806–834. https://doi.org/10.1037/0022-006X.46.4.806

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defence and two principles that warrant it. *Psychological Inquiry, 1*(2), 108-141. https://doi.org/10.1207/s15327965pli0102_1

Mid and South Essex Integrated Care System (2023, September 23). *Readers' panel*. https://www.midandsouthessex.ics.nhs.uk/get-involved/how/readers-panel/

Mid Cheshire Hospitals NHS Foundation Trust (2023, March 25). *Readers' panel*. https://www.mcht.nhs.uk/patients-and-visitors/readers-panel

Miller, A. C., Keenan, J. M., Betjemann, R. S., Willcutt, E. G., Pennington, B. F., & Olson, R. K. (2013). Reading comprehension in children with ADHD: Cognitive underpinnings of the centrality deficit. *Journal of Abnormal Child Psychology, 41*, 473-483. https://doi.org/10.1007/s10802-012-9686-8

Millis, K., Simon, S., & Tenbroek, N. S. (1998). Resource allocation during the rereading of scientific texts. *Memory & Cognition, 26*, 232–246. https://doi.org/10.3758/bf03201136

Mills, C. B., Diehl, V. A., Birkmire, D. P., & Mou, L. C. (1995). Reading procedural texts: Effects of purpose for reading and predictions of reading comprehension models. *Discourse Processes, 20*, 79–107. https://doi.org/10.1080/01638539509544932

Mo, L., Chen, H., Li, Y., Chen, Z., & He, X. (2007). Effects of event-related centrality on concept accessibility. *Discourse Processes, 43,* 229-254. https://doi.org/10.1080/01638530701226204

Moore, D., Lin, L., & Zabrucky, K. (2005). A source of metacomprehension inaccuracy. *Reading Psychology, 26*, 251–265.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., Wagenmakers, E. J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review, 23*, 103-123. https://doi.org/10.3758/s13423-015-0947-8

Morris, C. C. (1990). Retrieval processes underlying confidence in comprehension judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 223-232. https://doi.org/10.1037/0278-7393.16.2.223

Myers, M., & Paris, S. G. (1978). Children's metacognitive knowledge about reading. *Journal of Educational Psychology, 70*(5), 680–690. https://doi.org/10.1037/0022-0663.70.5.680

Narvaez, D., van den Broek, P., & Ruiz, A. B. (1999). The influence of reading purpose on inference generation and comprehension in reading. *Journal of Educational Psychology, 91*, 488-496. https://doi.org/10.1037/0022-0663.91.3.488

Nation, K., & Snowling, M. (1997). Assessing reading difficulties: The validity and utility of current measures of reading skill. *British Journal of Educational Psychology, 67*, 359-370. https://doi.org/10.1111/j.2044-8279.1997.tb01250.x

Navarro, D. J. (2021). If mathematical psychology did not exist we might need to invent it: A comment on theory building in psychology. *Perspectives on Psychological Science, 16*(4), 707–716. https://doi.org/10.1177/1745691620974769

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95*(1), 109–133. https://doi.org/10.1037/0033-2909.95.1.109

Nelson, T. O. (1986). ROC curves and measures of discrimination accuracy: A reply to Swets. *Psychological Bulletin, 100*(1), 128-132. https://doi.org/10.1037/0033-2909.100.1.128

Nelson, T. O. (1996). Gamma is a measure of the accuracy of predicting performance on one item relative to another item, not of the absolute performance on an individual item. *Applied Cognitive Psychology, 10*(3), 257–260. https://doi.org/10.1002/(SICI)1099-0720(199606)10:3<257::AID-ACP400>3.0.CO;2-9

Nelson, T. O., Dunlosky, J., Graf, A., & Narens, L. (1994). Utilization of metacognitive judgments in the allocation of study during multitrial learning. *Psychological Science, 5*(4), 207–213. https://doi.org/10.1111/j.1467-9280.1994.tb00502.x

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), Psychology of Learning and Motivation (Vol 26, pp. 125-173). Academic Press. https://doi.org/10.1016/S0079-7421(08)60053-5

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences, 236*, 333–380. https://doi.org/10.1098/rsta.1937.0005

Neyman, J., & Pearson, E. S. (1928a). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika, 20A*, 175-240. https://doi.org/10.2307/2331945

Neyman, J., & Pearson, E. S. (1928b). On the use and interpretation of certain test criteria for purposes of statistical inference: Part II. *Biometrika, 20A*, 263-294. https://doi.org/10.2307/2332112

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, 231*, 289-337. https://doi.org/10.1098/rsta.1933.0009

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2005). Metacognitive monitoring accuracy and student performance in the postsecondary classroom. *The Journal of Experimental Education, 74*, 7-28.

Nieuwland, M. S., & Van Berkum, J. J. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience, 18*, 1098-1111. https://doi.org/10.1162/jocn.2006.18.7.1098

Oakhill, J., Cain, K., & Elbro, C. (2014). *Understanding and teaching reading comprehension: A handbook*. Routledge.

Office for National Statistics. (2013). *2011 Census: Key statistics and quick statistics for Local Authorities in the United Kingdom*. https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentande mployeetypes/bulletins/keystatisticsandquickstatisticsforlocalauthoritiesintheunitedkingdom/2013-12-04/pdf

Omanson, R. C. (1980). The role centrality on story category saliency. *Learning Research and Development Center*. https://files.eric.ed.gov/fulltext/ED197324.pdf

O'Reilly, T., Sabatini, J., & Wang, Z. (2019). Using scenario-based assessments to measure deep learning. In K. Millis, D. Long, J. Magliano, & K. Wiemer (Eds.), *Deep comprehension: Multi-disciplinary approaches to understanding, enhancing, and measuring comprehension* (pp. 197-208). Routledge.

Otto, W., Barrett, T. C., Koenke, K. (1969). Assessment of children’s statements of the main idea in reading. In J.A. Figurel (Ed.), *Reading and realism* (pp. 692-697). International Reading Association.

Ozuru, Y., Best, R., Bell, C., Witherspoon, A., & McNamara, D. S. (2007). Influence of question format and text availability on the assessment of expository text comprehension. *Cognition and Instruction*, *25*(4), 399–438. https://doi.org/10.1080/07370000701632371

Ozuru, Y., Briner, S., Kurby, C. A., & McNamara, D. S. (2013). Comparing comprehension measured by multiple-choice and open-ended questions. *Canadian Journal of Experimental Psychology*, *67*, 215-227. https://doi.org/10.1037/a0032918

Ozuru, Y., Kurby, C. A., & McNamara, D. S. (2012). The effect of metacomprehension judgment task on comprehension monitoring and metacognitive accuracy. *Metacognition and Learning*, *7*(2), 113–131. https://doi.org/10.1007/s11409-012-9087-y

Paris, S. G., & Myers, M. (1981). Comprehension monitoring, memory, and study strategies of good and poor readers. *Journal of Reading Behavior*, *13*(1), 5–22. https://doi.org/10.1080/10862968109547390

Paulewicz, B., Siedlecka, M., & Koculak, M. (2020). Confounding in studies of metacognition: A preliminary causal analysis framework. *Frontiers in Psychology, 11*, 1-14. https://doi.org/10.3389/fpsyg.2020.01933

Perfetti, C. A. (1999). Cognitive research and the misconceptions of reading education. In J. Oakhill & R. Beard (Eds.), *Reading development and the teaching of reading: A psychological perspective* (pp. 42–58). Blackwell Science.

Perfetti, C. A., & Stafura, J. Z. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading, 18*(1), 22–37. https://doi.org/10.1080/10888438.2013.827687

Perfetti, C. A., & Stafura, J. Z. (2015). Comprehending implicit meanings in text without making inferences. In E. J. O'Brien, A. E. Cook, & R. F. Lorch, Jr. (Eds.), *Inferences during reading* (pp. 1–18). Cambridge University Press.

Perugini, M., Gallucci, M., & Costantini, G. (2018). A practical primer to power analysis for simple experimental designs. *International Review of Social Psychology, 31*(1), 1-23. https://doi.org/10.5334/irsp.181

Pilegard, C., & Mayer, R. E. (2015). Adding judgments of understanding to the metacognitive toolbox. *Learning and Individual Differences, 41*, 62–72. https://doi.org/10.1016/j.lindif.2015.07.002

Popper, K. R. (1963). *Conjectures and refutations: The growth of scientific knowledge*. Routledge.

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica, 104*, 1-15. https://doi.org/10.1016/S0001-6918(99)00050-5

Prinz, A., Golke, S., & Wittwer, J. (2020a). How accurately can learners discriminate their comprehension of texts? A comprehensive meta-analysis on relative metacomprehension accuracy and influencing factors. *Educational Research Review, 31*, 1-31. https://doi.org/10.1016/j.edurev.2020.100358

Prinz, A., Golke, S., & Wittwer, J. (2020b). To what extent do situation-model-approach interventions improve relative metacomprehension accuracy? Meta-analytic insights. *Educational Psychology Review, 32*(4), 917–949. https://doi.org/10.1007/s10648-020-09558-6

R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. https://www.R-project.org/

Ratajczak, M. (2020). *The effects of individual differences and linguistic features on reading comprehension of health-related texts* [Doctoral dissertation, Lancaster University]. Lancaster EPrints. https://doi.org/10.17635/lancaster/thesis/952

Ratzan, S. C. & Parker, R. M. (2000). Introduction. In C. R. Selden, M. Zorn, S. Ratzan, & R. M. Parker (Eds.) *Current bibliographies in medicine: Health literacy* (pp. v–vi). National Library of Medicine, U.S. Department of Health and Human Services.

Rausch, M., & Zehetleitner, M. (2017). Should metacognition be measured by logistic regression? *Consciousness and Cognition, 49*, 291-312. https://doi.org/10.1016/j.concog.2017.02.007

Rawson, K. A., & Dunlosky, J. (2002). Are performance predictions for text based on ease of processing? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(1), 69–80. https://doi.org/10.1037/0278-7393.28.1.69

Rawson, K. A., Dunlosky, J., & Thiede, K. W. (2000). The rereading effect: Metacomprehension accuracy improves across reading trials. *Memory & Cognition, 28*(6), 1004-1010. https://doi.org/10.3758/BF03209348

Rhodes, M. G., & Tauber, S. K. (2011). The influence of delaying judgments of learning (JOLs) on metacognitive accuracy: A meta-analytic review. *Psychological Bulletin, 137*, 131-148. https://doi.org/10.1037/a0021705

Rohrer, J. M., & Arslan, R. C. (2021). Precise answers to vague questions: Issues with interactions. *Advances in Methods and Practices in Psychological Science, 4*(2), 1-19. https://doi.org/10.1177/2515245921100736

Rothman, K. J., & Greenland, S. (2018). Planning study size based on precision rather than power. *Epidemiology, 29*(5), 599-603. https://doi.org/10.1097/EDE.0000000000000876.

Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin & Review, 26*, 452-467. https://doi.org/10.3758/s13423-018-1558-y

Sandberg, K., Bibby, B. M., & Overgaard, M. (2013). Measuring and testing awareness of emotional face expressions. *Consciousness and Cognition, 22*, 806-809. http://dx.doi.org/10.1016/j.concog.2013.04.015

Sandberg, K., Timmermans, B., Overgaard, M., & Cleeremans, A. (2010). Measuring consciousness: Is one measure better than the other? *Consciousness and Cognition, 19*, 1069-1078. https://doi.org/10.1016/j.concog.2009.12.013

Scheck, P., & Nelson, T. O. (2005). Lack of pervasiveness of the underconfidence-with-practice effect: Boundary conditions and an explanation via anchoring. *Journal of Experimental Psychology: General, 134*, 124-128. https://doi.org/10.1037/0096-3445.134.1.124

Schommer, M., & Surber, J. R. (1986). Comprehension-monitoring failure in skilled adult readers. *Journal of Educational Psychology, 78*(5), 353–357. https://doi.org/10.1037/0022-0663.78.5.353

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality, 47*(5), 609–612. https://doi.org/10.1016/j.jrp.2013.05.009

Schraw, G. (1994). The effect of metacognitive knowledge on local and global monitoring. *Contemporary Educational Psychology, 19,* 143-154. https://doi.org/10.1006/ceps.1994.1013

Schraw, G. (2009a). A conceptual analysis of five measures of metacognitive monitoring. Metacognition and Learning, 4(1), 33–45. https://doi.org/10.1007/s11409-008-9031-3

Schraw, G. (2009b). Measuring metacognitive judgments. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 415–429). Routledge/Taylor & Francis Group.

Schraw, G., Kuch, F., & Gutierrez, A. P. (2013). Measure for measure: Calibrating ten commonly used calibration scores. *Learning and Instruction, 24,* 48-57. http://dx.doi.org/10.1016/j.learninstruc.2012.08.007

Schroeder, S. (2011). What readers have and do: Effects of students' verbal ability and reading time components on comprehension with and without text availability. *Journal of Educational Psychology, 103*(4), 877–896. https://doi.org/10.1037/a0023731

Schwartz, B. L. (1994). Sources of information in metamemory: Judgements of learning and feelings of knowing. *Psychonomic Bulletin & Review, 1*(3), 357-375. https://doi.org/10.3758/BF03213977

Schwartz, B. L., & Metcalfe, J. (1994). Methodological problems and pitfalls in the study of human metacognition. In J. Metcalfe, & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 93-114). The MIT Press. https://doi.org/10.7551/mitpress/4561.003.0007

Serra, M. J., & Dunlosky, J. (2010). Metacomprehension judgements reflect the belief that diagrams improve learning from text. *Memory, 18*(7), 698–711. https://doi.org/10.1080/09658211.2010.506441

Serra, M. J., & Metcalfe, J. (2009). Effective implementation of metacognition. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), Handbook of metacognition in education (pp. 278-298). Routledge/Taylor & Francis Group.

Shanks, L. L., & Serra, M. J. (2014). Domain familiarity as a cue for judgments of learning. *Psychonomic Bulletin & Review, 21*(2), 445–453. https://doi.org/10.3758/s13423-013-0513-1

Shipley, W. C. (1940). A self-administering scale for measuring intellectual impairment and deterioration. *The Journal of Psychology: Interdisciplinary and Applied, 9*, 371-377. https://doi.org/10.1080/00223980.1940.9917704

Shipley, W. C., Gruber, C. P., Martin, T. A., & Klein, A. M. (2009). *Shipley Institute of Living Scale-2*. Los Angeles, CA Western Psychological Services.

Sivula, T., Magnusson, M., Matamoros, A. A., & Vehtari, A. (2020). Uncertainty in Bayesian leave-one-out cross-validation based model comparison. *arXiv:2008.10296*. https://doi.org/10.48550/arXiv.2008.10296

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science, 3*, 1-17. https://doi.org/10.1098/rsos.160384

Snijders, T. A., & Bosker, R. J. (2011). *Multilevel analysis* (2nd Edition). Sage.

Somerset NHS Clinical Commissioning Group (2019). *Communications and engagement strategy 2019-2022*. https://www.somersetccg.nhs.uk/wp-content/uploads/2020/06/Communications-and-engagement-strategy-and-action-plan-approved-September-2019.pdf

Sørensen, K., Pelikan, J. M., Röthlin, F., Ganahl, K., Slonska, Z., Doyle, G., Fullam, J. Kondilis, B., Agrafiotis, D., Uiters, E., Falcon, M., Mensing, M., Tchamov, K., van den Broucke, S., & Brand, H. (2015). Health literacy in Europe: Comparative results of the European health literacy survey (HLS-EU). *European Journal of Public Health, 25*(6), 1053-1058. https://doi.org/10.1093/eurpub/ckv043

Stafura, J. Z., & Perfetti, C. A. (2014). Word-to-text integration: Message level and lexical level influences in ERPs. *Neuropsychologia, 64*, 41-53. https://doi.org/10.1016/j.neuropsychologia.2014.09.012

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician, 49*(1), 108-112. https://doi.org/ 10.1080/00031305.1995.10476125

Tauber, S. (U.) K., & Dunlosky, J. (2016). A brief history of metamemory research and handbook overview. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford handbook of metamemory* (pp. 7–21). Oxford University Press

Thiede, K. W., & Anderson, M. C. M. (2003). Summarizing can improve metacomprehension accuracy. *Contemporary Educational Psychology, 28*(2), 129–160. https://doi.org/10.1016/S0361-476X(02)00011-5

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*(1), 66–73. https://doi.org/10.1037/0022-0663.95.1.66

Thiede, K. W., Griffin, T. D., Wiley, J., & Anderson, M. C. M. (2010). Poor metacomprehension accuracy as a result of inappropriate cue use. *Discourse Processes, 47*(4), 331-362. https://doi.org/10.1080/01638530902959927

Thiede, K. W., Griffin, T. D., Wiley, J., & Redford, J. S. (2009). Metacognitive monitoring during and after reading. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 85-106). Routledge.

Thiede, K. W., Wiley, J., & Griffin, T. D. (2011). Test expectancy affects metacomprehension accuracy. *British Journal of Educational Psychology, 81*, 264-273. https://doi.org/10.1348/135910710X510494

Thiede, K. W., Wright, K. L., Hagenah, S., & Wenner, J. (2019). Drawings as diagnostic cues for metacomprehension judgement. In N. Feza (Ed.), *Metacognition in learning* (pp. 65-80). IntechOpen. https://doi.org/10.5772/intechopen.78892

Thiede, K. W., Wright, K. L., Hagenah, S., Wenner, J., Abbott, J., & Arechiga, A. (2022). Drawing to improve metacomprehension accuracy. *Learning and Instruction, 77*, 101541. https://doi.org/10.1016/j.learninstruc.2021.101541

Thorndike, E. L. (1917). Reading as reasoning: A study of mistakes in paragraph reading. *Journal of Educational Psychology, 8*(6), 323–332. https://doi.org/10.1037/h0075325

Thorndyke, P. W. (1977). Cognitive structures in comprehension and memory of narrative discourse. *Cognitive Psychology, 9*, 77-110. https://doi.org/10.1016/0010-0285(77)90005-6

Tobias, S., & Everson, H. T. (2009). The importance of knowing what you know: A knowledge monitoring framework for studying metacognition in education. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 107-127). Routledge/Taylor & Francis Group.

Trabasso, T., Secco, T., & van den Broek, P. W. (1984). Causal cohesion and story coherence. In H. Mandl, N. L. Stein, & T. Trabasso (Eds.*), Learning and comprehension of text* (pp. 83-111). Lawrence Erlbaum.

Trabasso, T., & Sperry, L. L. (1985). Causal relatedness and importance of story events. *Journal of Memory and Language, 24,* 595-611. https://doi.org/10.1016/0749-596X(85)90048-8

Trabasso, T., & van den Broek, P. W. (1985). Causal thinking and the representation of narrative events. *Journal of Memory and Language, 24*, 612-630. https://doi.org/10.1016/0749-596X(85)90049-X

Trafimow, D. (2015). A defence against the alleged unreliability of difference scores. *Cogent Mathematics*, *2*(1), 1064626. https://doi.org/10.1080/23311835.2015.1064626

Trafimow, D. (2019) Five nonobvious changes in editorial practice for editors and reviewers to consider when evaluating submissions in a post $p < 0.05$ universe. *The American Statistician, 73,* 340-345. https://doi.org/10.1080/00031305.2018.1537888

Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131. https://doi.org/10.1126/science.185.4157.1124

Tzeng, Y., van den Broek, P., Kendeou, P., & Lee, C. (2005). The computational implementation of the landscape model: Modeling inferential processes and memory representations of text comprehension. *Behavior Research Methods, 37*(2), 277–286. https://doi.org/10.3758/BF03192695

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., & Yau, C. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers, 1,* 1-26. https://doi.org/10.1038/s43586-020-00001-2

van den Broek, P. (1988). The effects of causal relations and hierarchical position on the importance of story statements. *Journal of Memory and Language, 27,* 1–22. https://doi.org/10.1016/0749-596X(88)90045-9 van den Broek, P., Bohn-Gettler, C. M., Kendeou, P., Carlson, S., & White, M. J. (2011). When a reader meets a text: The role of standards of coherence in reading comprehension. In M. T. McCrudden, J. P. Magliano, & G. Schraw (Eds.), *Text relevance and learning from text* (pp. 123-139). IAP Information Age Publishing.

van den Broek, P., & Helder, A. (2017). Cognitive processes in discourse comprehension: Passive processes, reader-initiated processes, and evolving mental representations. *Discourse Processes, 54*(5-6), 360–372. https://doi.org/10.1080/0163853X.2017.1306677

van den Broek, P., Helder, A., & Van Leijenhorst, L. (2013). Sensitivity to structure centrality. In A. Britt, S. Goldman, & J. F. Rouet (Eds.), *Reading – From words to multiple texts* (pp. 132-146). Routledge.

van den Broek, P., Lorch, E. P., & Thurlow, R. (1996). Children's and adult's memory for television stories: The role of causal factors, story-grammar categories, and hierarchical level. *Child Development, 67*, 3010-3028. https://doi.org/10.2307/1131764

van den Broek, P., Lorch, R. F., Linderholm, T., & Gustafson, M. (2001). The effects of readers' goals on inference generation and memory for texts. *Memory & Cognition, 29*, 1081-1087. https://doi.org/10.3758/BF03206376

van den Broek, P., Risden, K., & Husebye-Hartman, E. (1995). The role of reader's standards of coherence in generation of inferences during reading. In E. P. Lorch & E. J. O'Brien (Eds.), *Sources of Coherence in Reading* (pp. 353–374). Erlbaum.

van den Broek, P., & Trabasso, T. (1986). Causal networks versus goal hierarchies in summarizing text. *Discourse Processes, 9*, 1–15. https://doi.org/10.1080/01638538609544628 van den Broek, P., Young, M., Tzeng, Y., & Linderholm, T. (1999). The Landscape model of reading: Inferences and the online construction of memory representation. In H. van Oostendorp & S. R. Goldman (Eds.), *The construction of mental representations during reading* (pp. 71–98). Lawrence Erlbaum Associates Publishers.

van der Ark, A., & van Aert, R. C. M. (2014). Comparing confidence intervals for Goodman and Kruskal's gamma coefficient. *Journal of Statistical Computation and Simulation, 85*(12), 1-15. https://doi.org/ 10.1080/00949655.2014.932791

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. Academic Press

Vankov, I., Bowers, J., & Munafo, M.R. (2014). On the persistence of low power in psychological science. *The Quarterly Journal of Experimental Psychology, 67*, 1037-1040. https://doi.org/10.1080/17470218.2014.885986

Vehtari, A., Gelman, A., & Gabry J. (2016). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing, 27*, 1413-1423. doi:10.1007/s11222-016-9696-4.

Verde, M. F., Macmillan, N. A., & Rotello, C. M. (2006). Measures of sensitivity based on a single hit rate and false alarm rate: The accuracy, precision, and robustness of *d'*, A_z, and *A*'. *Perception & Psychophysics*, *68*, 643-654. https://doi.org/10.3758/BF03208765

Vössing, J., & Stamov-Roßnagel, C. (2016). Boosting metacomprehension accuracy in computer-supported learning: The role of judgment task and judgment scope. *Computers in Human Behaviour, 54*, 73-82. https://doi.org/10.1016/j.chb.2015.07.066

Vuorre, M., & Metcalfe, J. (2022). Measures of relative metacognitive accuracy are confounded with task performance in tasks that permit guessing. *Metacognition and Learning, 17*, 269-291. https://doi.org/10.1007/s11409-020-09257-1

Walther, R. A., & Moore, J. L. (2005). The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography, 28*, 815-829. https://doi.org/10.1111/j.2005.0906-7590.04112.x

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *The Journal of Machine Learning Research, 11*, 3571-3594.

Weaver, C. A. III (1990). Constraining factors in calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(2), 214–222. https://doi.org/10.1037/0278-7393.16.2.214

Weaver, C. A. III, & Bryant, D. S. (1995). Monitoring of comprehension: The role of text difficulty in metamemory for narrative and expository text. *Memory & Cognition, 23*, 12-22. https://doi.org/10.3758/BF03210553

Weaver, C. A. III, Bryant, D. S., & Burns, K. D. (1995). Comprehension monitoring: Extensions of the Kintsch and van Dijk model. In C. A. Weaver III, S. Mannes, & C. R. Fletcher (Eds.), *Discourse comprehension: Essays in honor of Walter Kintsch* (pp. 177-193). Lawrence Erlbaum Associates, Inc.

Weisberg, D. S., Keil, F. C., Goodstein, J., Rawson, E., & Gray, J. R. (2008). The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience, 20*(3), 470-477. https://doi.org/10.1162/jocn.2008.20040

Weiss, B. D., Hart, G., McGee, D. L., & D'Estelle, S. (1992). Health status of illiterate adults: Relation between literacy and health status among persons with low literacy skills. *The Journal of the American Board of Family Practice, 5*(3), 257-264. https://doi.org/10.3122/jabfm.5.3.257

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology, 7,* 1-12. https://doi.org/10.3389/fpsyg.2016.01832

Wiley, J. (2019). Picture this! Effects of photographs, diagrams, animations, and sketching on learning and beliefs about learning from a geoscience text. *Applied Cognitive Psychology, 33*(1), 9–19. https://doi.org/10.1002/acp.3495

Wiley, J., Griffin, T. D., Jaeger, A. J., Jarosz, A. F., Cushen, P. J., & Thiede, K. W. (2016). Improving metacomprehension accuracy in an undergraduate course context. *Journal of Experimental Psychology: Applied, 22*(4), 393–405. https://doi.org/10.1037/xap0000096

Wiley, J., Griffin, T. D., & Thiede, K. W. (2005). Putting the comprehension in metacomprehension. *Journal of General Psychology, 132*(4), 408-428. https://doi.org/10.3200/GENP.132.4.408-428

Wiley, J., Griffin, T. D., & Thiede, K. W. (2008). To understand your understanding, you must understand what understanding means. *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Wiley, J., Jaeger, A. J., Taylor, A. R., & Griffin, T. D. (2018). When analogies harm: The effects of analogies on metacomprehension. *Learning and Instruction, 55,* 113–123. https://doi.org/10.1016/j.learninstruc.2017.10.001

Winograd, P., & Johnston, P. (1980). Comprehension monitoring and the error detection paradigm. *Journal of Literacy Research, 14*(1), 61-76. https://doi.org/10.1080/10862968209547435

Wixted, J. T. (2020). The forgotten history of Signal Detection Theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46*(2), 201-233. http://dx.doi.org/10.1037/xlm0000732

Yang, C., Sun, B., & Shanks, D. R. (2018). The anchoring effect in metamemory monitoring. *Memory & Cognition, 46,* 384-397. https://doi.org/10.3758/s13421-017-0772-6

Yang, C., Zhao, W., Yuan, B., Luo, L., & Shanks, D. R. (2022). Mind the gap between comprehension and metacomprehension: Meta-analysis of metacomprehension accuracy and intervention effectiveness. *Review of Educational Research, 0*, 1-52. https://doi.org/10.3102/00346543221094083

Yarkoni, T. (2009). Big correlations in little studies: Inflated fMRI correlations reflect low statistical power – commentary on Vul et al. (2009). *Perspectives on Psychological Science, 4*(3), 294-298. https://doi.org/10.1111/j.1745-6924.2009.01127

Yeari, M., & Lantin, S. (2020). The origin of centrality deficit in text memory and comprehension by poor comprehenders: A think-aloud study. *Reading and Writing, 34*, 595-625. https://doi.org/10.1007/s11145-020-10083-9

Yeari, M., Oudega, M., & van den Broek, P. (2017). The effect of highlighting on processing and memory of central and peripheral text information: Evidence from eye movements. *Journal of Research in Reading, 40*, 365-383. https://doi.org/10.1111/1467-9817.12072

Yeari, M., van den Broek, P., & Oudega, M. (2015). Processing and memory of central versus peripheral information as a function of reading goals: Evidence from eye-movements. *Reading and Writing, 28,* 1071-1097. https://doi.org/10.1007/s11145-015-9561-4

Yussen, S. R., & Smith, M. C. (1990). Detecting general and specific errors in expository texts. *Contemporary Educational Psychology, 15*, 224-240. https://doi.org/10.1016/0361-476X(90)90020-2

Zabrucky, K. M. (2010). Knowing what we know and do not know: Educational and real world implications. *Procedia – Social and Behavioral Sciences, 2*(2), 1266-1269. https://doi.org/10.1016/j.sbspro.2010.03.185

Zabrucky, K., Moore, D., & Schultz, N. R. (1987). Evaluation of comprehension in young and old adults. *Developmental Psychology, 23*(1), 39–43. https://doi.org/10.1037/0012-1649.23.1.39

Zabrucky, K., Moore, D., & Schultz, N. R. (1993). Young and old adults' ability to use different standards to evaluate understanding. *Journal of Gerontology, 48*(5), P238–P244. https://doi.org/10.1093/geronj/48.5.P238

Zhao, Q. (2022). Absolute standing feedback is more influential than relative standing feedback. *Journal of Educational Psychology, 114*, 701-715. https://doi.org/10.1037/edu0000676

Zhao, Q., & Linderholm, T. (2008). Adult metacomprehension: Judgment processes and accuracy constraints. *Educational Psychology Review, 20*, 191-206. https://doi.org/10.1007/s10648-008-9073-8

Zwaan, R. A. (2016). Situation models, mental simulations, and abstract concepts in discourse comprehension. *Psychonomic Bulletin & Review, 23*, 1028-1034. https://doi.org/10.3758/s13423-015-0864-x

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*, 162-185. https://doi.org/ 10.1037/0033-2909.123.2.162