Investigating the construct measured by banked gap-fill items: Evidence from eye-tracking

Abstract

This study investigates test-takers' processing while completing banked gap-fill tasks, designed to test reading proficiency, in order to test theoretically based expectations about the variation in cognitive processes of test-takers across levels of performance. Twenty-eight test-takers' eye traces on 24 banked gap-fill items (on six tasks) were analysed according to seven on-line eye-tracking measures representing overall, text and task processing. Variation in processing was related to test-takers' level of performance on the tasks overall. In particular, as hypothesised, lower-scoring students exerted more cognitive effort on local reading and lower-level cognitive processing in contrast to test-takers who attained higher scores. The findings of different cognitive processes associated with variation in scores illuminate the construct measured by banked gap-fill items, and therefore have implications for test design and the validity of score interpretations.

Keywords

Testing reading, Eye-tracking, Banked gap-fill, Cloze, Cognitive processing

Introduction

Banked gap-fill items belong to a family of item types (sometimes called *gap-fill*, sometimes *cloze*) in which portions of text, typically individual words, are removed from a text and test-takers are asked to reconstruct those missing words. Banked gap-fill tasks, in particular, are often used as part of reading tests (e.g. in the Aptis test (British Council, n.d.), PTE Academic test (Pearson, n.d.)). Given this fact, it is surprising that banked gap-fill items are seemingly under-researched and it is not necessarily clear exactly what is being measured by this member of the item type family despite the considerable interest in such test items. For example, Oller and Jonz (1994a) suggested that various types of knowledge are required for successful completion of this item family. Their suggestion of knowledge required includes phonological, semantic, and syntactic knowledge of the target language alongside general knowledge of the world and inferences that can be made from world knowledge. Such broad hypotheses proved difficult to investigate leaving questions in the field about the meaning of scores obtained through tests using such items.

This study aims to investigate the construct of banked gap-fill items by examining the cognitive processing of study participants during the on-line completion of this specific item type, as measured by an eye-tracker. The study will examine the types of processing undertaken by higher- and lower-performing participants in order to better delineate the construct measured. Initially though, a definition of the item type under scrutiny – banked gap-fill items – will be given and the item type will be contextualised within its broader item-type family, in order to clarify the focus of this study.

Literature Review

The broader item-type family – whether referred to as gap-fill or cloze – comprises tasks which consist of a text from which a number of words have been deleted and for which the

test-taker is required to restore the omitted words (Davies et al., 1999). In some cases, the words are deleted such that every n^{th} word is omitted – typically somewhere between every 4^{th} word and every 12^{th} word. This format is often called *fixed-ratio cloze* (Oller & Jonz, 1994a). In other cases, words are deleted rationally according to some specific purpose (e.g., prepositions to test for preposition usage knowledge). Some authors use the term *gap-fill* to specifically refer to the latter (e.g, Alderson, 2000), others uses the term *rational-deletion cloze* (e.g., Oller & Jonz, 1994a). Henceforth, we will use the term *cloze* when referring to fixed-ratio cloze, *gap-fill* when referring to rational-deletion cloze, and *gap-filling* as the umbrella term for the entire family of item types.

Gap-filling tasks require the test-taker to place into a blank space in the text a word that the test developer deleted. The test-taker may need to complete the text by selecting a word from a number of options provided for each individual gap. This gap-filling format is coined as a form of *multiple-choice* (Jonz, 1976). Alternatively, the test-takers themselves may be required to generate lexical items to fill the gaps. These are typically referred to as *open* gap-filling tasks (Alderson & Cseresznyés, 2003). In *banked* gap-filling tasks, words omitted from the text are provided, but randomly ordered in a 'bank' outside the text, often with additional words as distractors. For each gap in the text the test-taker needs to select a word from the bank to reconstruct the passage (Alderson & Cseresznyés, 2003). It is likely that the method of replacing the words in the text has an influence on the processing of the task; for example, in multiple-choice or banked gap-filling items deductive reasoning can be used to select the best fit from among the alternatives, whereas in open gap-filling this is not possible.

Oller and Jonz (1994b) ground the construct of gap-filling items (note that they use the term 'cloze') in terms of *coherence*; specifically, the extent to which the texts or their interpretations actually match with experiences in the 'real' world outside the text (Oller &

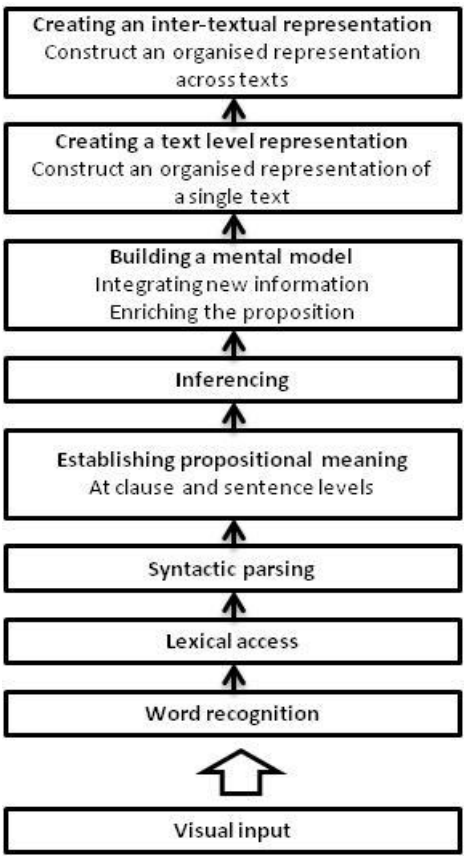
Jonz, 1994). They suggest a tri-partite and interactive model of the sources of score variance of these types of items where, all other things being held constant: 1) more coherent texts will elicit higher scores, 2) more proficient (i.e. experienced, knowledgeable, intelligent, motivated) test-takers will score higher, and 3) more proficient producers (item writers) of coherent text will generate texts which will elicit higher scores. However, Oller & Jonz (1994c) admit that "the sources of variability can interact in complex ways" (p.46). Therefore, more information about what exactly drives the relationship between the variance in test-scores and the cognitive processing stimulated by the test tasks is needed to understand the construct that the test measures from a processing perspective. Potentially, this type of information can be extracted by examining the online behaviours of test-takers while responding to items.

Online behaviours of test-takers in gap-filling tests have not been investigated extensively despite the long history of gap-filling research. The question of whether gap-filling items can be used as valid measures of reading proficiency, i.e. the ability to "receiv[e] and interpret information encoded in language via the medium of [online] print" (Urquhart & Weir, 1998, p.22), has been investigated. Fixed-ratio cloze tasks were originally produced by Taylor (1953) in order to assess text readability for first language readers, and in the 1970s they also emerged as a test format which, it was claimed, could be used to assess both L2 reading and general language proficiency (Alderson, 2000). Some researchers, however, have argued that the construct measured by a cloze item is closely linked to the specific word deleted (Alderson, 1980; Bachman, 1985; Jonz, 1990). Miller and Coleman (1967), for example, found that changing the starting point for deleting every 5th word of a cloze text changed test results. Alderson (2000) similarly illustrated, for rational-deleted gap-fill tasks, how different versions, with different words deleted but based on the same text, can measure different sets of skills and knowledge dependent on the words deleted. He thus argues that the

rational deletion of words in a text gives the test constructor more control over the construct of their test. However, as Yamashita (2003, p. 269) points out, even if a rational-deletion gap-fill procedure is used there is no guarantee that the omitted words will actually test what the test constructors intend.

A specific controversy surrounding gap-filling tasks is whether they can measure more global reading skills, or whether they just assess more local and so-called lower-level reading processes (which may shift the construct being measured more towards language-in-use tasks than reading tasks). In a recent and increasingly influential model of reading comprehension, Khalifa and Weir (2009) hierarchically list the cognitive processes in reading in the central processing core of their model as, from lowest to highest level: word recognition (word decoding), lexical access (extracting meaning), syntactic parsing (deciphering grammatical information in the text), and establishing propositional meaning at the clause/sentence level – i.e. lower-order processes; and inferencing, building a mental model of the text by integrating various propositions, creating a text-level representation by forming a discourse-level structure of the text as a whole, and creating an intertextual representation by integrating information from multiple texts – i.e. higher-order processes. Figure 1 shows the central processing core visually.

Figure 1: The central processing core of the Khalifa and Weir's (2009) cognitive model of reading (Khalifa & Weir, 2009, p.43).



Khalifa and Weir's model (2009) posits that accurate and automatic word recognition, the lowest level process, is key to efficient reading comprehension (see also e.g. Grabe, 2009; Perfetti, 1985; Wagner & Stanovich, 1996). Automaticity is said to be the result of repeated experience, meaning that higher-performing readers, who have more efficient word recognition processes, have more attentional capacity available in their working memory to divert to higher-level reading processes. The converse is true for lower-performing readers who are theorised to have less attentional capacity to invest in higher-level processing (Khalifa & Weir, 2009).

Expanding on this, the processes in Khalifa and Weir's model (2009) build upon one another, from the bottom-up perspective, i.e. words being parsed into phrases, phrases into sentences, etc., and disruptions at the word recognition level can cascade upwards. Under time pressure, such as when sitting a test, this would mean that the individual who has less

efficient word recognition abilities faces greater difficulty in parsing phrases and sentences, and thus a greater proportion of their time-limited cognitive capacity will be expended on these lower-level processes. This is not to suggest, however, that bottom-up processing is the only method of extracting meaning from a text. Khalifa and Weir (2009) suggest that top-down processing (e.g., using background information to aid in the comprehension of lower-level units of meaning) also plays a part in reading and that there are two distinct uses for it; firstly, to enrich understanding of the text, and secondly, to aid in decoding when lower-level processing abilities are inadequate. Readers with poor lower-level processing ability are posited to be more likely to be engaging in the second use (to aid decoding), to the detriment of the first use, meaning they have less available capacity for enriching a more global understanding of the text (Khalifa & Weir, 2009).

The findings on which reading processes can actually be assessed by gap-filling tasks are mixed. With reference to fixed-ratio cloze tasks, numerous studies support the position that these can measure higher-level reading processes (Bachman, 1985; Brown, 1983; Chihara, Oller, Weaver, & Chavez-Oller, 1977; Cziko, 1978; Gamarra & Jonz, 1987; Jonz, 1987, 1990; McKenna & Layton, 1990; Oller, 1975; Taylor, 1957), while other research supports the argument that they are poor measures of higher-level reading processes (Alderson, 1980; Kibby, 1980; Klein-Braley, 1983; Leys, Fielding, Herman, & Pearson, 1983; Markham, 1985; Porter, 1978, 1983; Shanahan & Kamil, 1983). A potential explanation for these conflicting results, posited by Brown (2004, p. 84), may be found in the reading ability level of the test-takers in relation to the difficulty of the items. Brown asserts that with lower proficiency test-takers, these tasks are likely testing more at the intra-sentential, lower processing level, as these test-takers cannot handle more complex textual information, whereas for higher proficiency test-takers, they may test both intra-sentential and inter-sentential processing, as these test-takers have the linguistic ability to make use of

more global textual information. Brown's proposition might be elucidated more formally in the cognitive frame of the Khalifa and Weir (2009) model. Specifically, under time pressure, lower-performing readers will not have as much surplus attention capacity to divert to higher-level, inter-sentential reading processes as will more skilled readers.

Until relatively recently, and in the majority of the abovementioned studies, most research investigating the above issues has been product-oriented. Namely, test-takers' scores on gap-filling tasks have been analysed in terms of other relevant variables. Increasingly, however, the investigation of the processes of test-taking forms a critical component of the test validation process (Weir, 2005); as Alderson (2000) asserts: "The validity of the test relates to the interpretation of the correct responses to items, so what matters is not what the test constructors believe an item to be testing, but which responses are considered correct, and what process underlies them" (p.97). This concern about cognitive processes in investigations of validity suggests the need to assess the extent to which items elicit the same cognitive processes as readers perform when reading outside the testing context, i.e. in "real life". Khalifa and Weir (2009) claim their framework can be used as a "suitable model of real-life reading which can then be applied to evaluate the reading tasks employed in tests" (p.41). An investigation of the cognitive processes undertaken by test-takers when responding to a particular item type, with regards to performance on that item type, is one strategy that can help us to more clearly define the construct it measures. Currently, however, relatively few studies have analysed the processes test-takers undertake when responding to gap-filling items, and even fewer for banked gap-fill items.

Storey (1997) administered 13 multiple-choice gap-fill items designed to assess discourse processing strategies to 25 Hong-Kong Chinese test-takers. Insights into test-takers' processing during task completion were gained via concurrent introspection and immediate retrospection. In the analysis, the items were divided into the categories *discourse*

markers and *cohesive devices* according to the type of word deleted from the text. Storey found that items composed of deleted discourse markers elicited more cross-sentential reading behaviour, encouraging the test-takers to deconstruct the rhetorical structure of the text. Conversely, items targeting cohesive devices elicited more local reading behaviour.

In another study, Sasaki (2000) undertook a partial, process-oriented replication of an earlier product-focussed study by Chihara et al. (1977), exploring the effect of cultural familiarity on a cloze task. To this end, Sasaki used one text, but replaced a number of words to create a culturally familiar adaptation of the same text. Immediate retrospections by 60 Japanese learners of English indicated that test-takers who responded to the culturally familiar text used more within-sentence and within-clause information than those who completed the unfamiliar text.

In an exploratory study of test-takers' cognitive processing on open gap-fill items, Yamashita (2003) found some evidence in support of Brown's (2004) hypothesis, mentioned above, though her sample size was small. Yamashita compared the concurrent verbal protocols of six less-skilled with six skilled Japanese readers (selected on the basis of reading test scores) on a 16-item gap-fill task. The study revealed that less skilled readers reported more emphasis on local grammatical information, whereas more skilled readers tended to use local information in the text more to confirm their answers, which they had already established through more global reading.

The only process-oriented study investigating banked gap-fill tasks, to our knowledge, is by Gao and Gu (2008) who explored the task completion process of three groups of Chinese English learners (six each in a high, medium, and low group based on CET-4 reading scores) on a 10-item banked gap-fill task. Verbal protocol data from their study indicated that the most common response strategy was to refer to local (clause level) information as

opposed to sentential, textual or inter-textual information. This contrasted with Yamashita (2003) who found that the most commonly reported strategy was to extract information from the text as a whole. The salient difference between these two studies potentially relates to the difference in item type, i.e. presence of an option bank in Gao and Gu (2008) versus no options provided whatsoever in Yamashita (2003). In fact, Gao and Gu (2008, p. 8) mention that many test-takers reported that matching the words from the bank with the correct part of speech was a faster method to answer the questions.

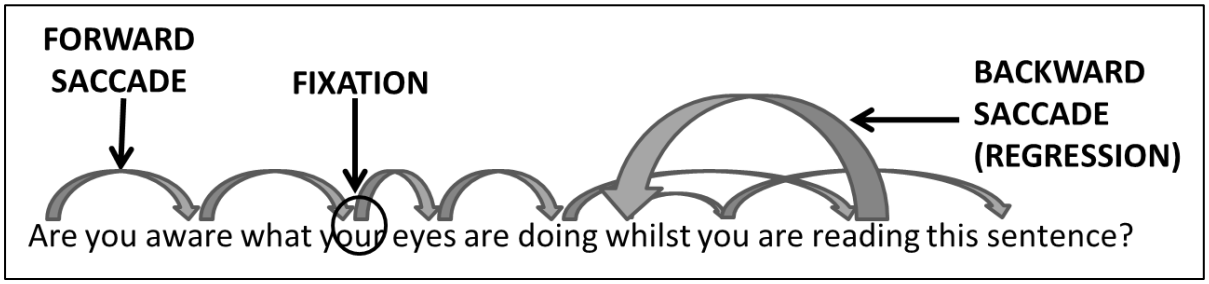
While these initial exploratory studies provide some interesting insights into what gap-filling tasks are actually testing, they do have limitations. First, the findings of the above studies are based on data collected with one gap-filling text only. Second, the populations in each study had a homogeneous cultural, educational and L1 background. These first two features have implications for the generalisability of the studies. Third, it has been suggested that the difference between the rational-deletion and fixed-ratio procedure and also between variations of each type may mean that the different gap-filling task types measure different constructs (Chapelle & Abraham, 1990; Soudek & Soudek, 1983). Thus, the conclusions drawn in previous research may not fully carry over to the banked gap-fill item type under scrutiny here (with the exception of Gao & Gu, 2008); the presence of the word bank may lead to differing item processing patterns. Fourth, as Yamashita (2003) observed, when responses to items were highly automatised in the test-taker, the verbal report data could not provide processing information. In fact, Yamashita doubted the extent to which the verbal protocols could reveal information about participants' lower-level cognitive processing. It is thus plausible that only partial information has been captured in these studies. Furthermore, verbal protocol methods have been criticised in that they risk veridicality and reactivity (Bowles, 2010), i.e. inaccurate reflections of one's thought processes and alterations of the thought process due to the activity of talking out loud, respectively.

One possibility for gathering data on the cognitive processing during reading which does not rely on the perception and recall of processes is via eye-tracking. Although eye-tracking is a relatively new method to the field of language testing, it seems particularly valuable as a data collection instrument due to the minimal cognitive disruption of the test-taking process, since test-takers complete the tasks as they would normally (especially with non-intrusive eye-trackers such as Tobii-300, used in this study, which requires no restraint apparatus) and without any additional processing requirements such as talk-alouds. Additionally, eye-tracking has been shown to generate unique insights into the testing of reading (see e.g. Bax, 2013; Bax & Weir, 2012; Brunfaut & McCray, 2015). Furthermore, Brunfaut & McCray (2015), for example, found that eye-tracking gave proportionally more insights into lower-level processing than did a verbal protocol method (stimulated recall). Thus, eye-tracking may usefully address some key weaknesses of verbal protocols.

The use of eye-tracking in reading research rests on the assumption that there is a close link between the point in a text on which our eyes fixate and the focus of our attention at that moment (Rayner, 1998). When we read, our eyes do not move continuously across the line of text (Rayner, Juhasz, & Pollatsek, 2007; Rayner, Pollatsek, Ashby, & Clifton, 2012). Rather, they make a series of small jumps along the line, which are termed 'saccades'. As our visual field varies in the level of detail our eyes can capture, i.e. its 'acuity' (Rayner et al., 2007; Rayner, 1998), the saccades are necessary to bring relatively small dense text into the region of highest acuity for processing. At the end of each saccade, our eyes pause on a point on the page, often for just a fraction of a second. These pauses are termed 'fixations' and it is during these fixations that a portion of the text is extracted for processing. A visual representation is provided in Figure 2 (adapted from Brunfaut & McCray, 2015, p.10). Given the potential of eye-tracking to provide a unique perspective on moment-to-moment

processing generated when completing test items, the decision was made to utilise it in this study.

Figure 2: Conceptual illustration of eye movements while reading



The present study

The primary aim of the present study was to gain insights into the construct measured by the banked gap-fill item type when used to test reading. More specifically, the study investigates the extent to which Brown's (2004) hypothesis applies to the banked gap-fill format, namely that the employment of higher-level reading processes when completing this task type may depend on test-taker performance level. For the purposes of our study, the reading processes were defined as those specified in the central processing core of Khalifa and Weir's (2009) cognitive processing model of reading comprehension, mentioned above. This model was chosen because it is increasingly utilised in the context of reading testing research (e.g., Brunfaut & McCray, 2015; Ilc & Stopar, 2015; Weir, Hawkey, Green, & Devi, 2012). It is felt that the model provides a valuable heuristic for the conceptualisation of cognitive processing during reading, in the context of language testing, and that it helps bridge the gap purported to exist between current theories of reading comprehension and the manner in which reading is tested (Anderson, Bachman, Perkins, & Cohen, 1991; Engelhard, 2001; Pollitt & Taylor, 2006). Furthermore, empirical support for the processes described by the model has been found in recent reading test research (see e.g. Brunfaut & McCray, 2015; Owen, 2015).

A second aim of the study was to address some of the methodological limitations of previous process-oriented studies, as discussed above. Therefore, a heterogeneous test-taking population and a larger sample of gap-filling tasks were sought. Furthermore, a method other than verbal reports was adopted, i.e. eye-tracking.

On the basis of these aims, i.e. to investigate test-takers' processing while completing banked gap-fill tasks using a promising alternative and currently under-researched methodology and to explore the relationship between processing and performance level, seven hypotheses were formulated on processing characteristics as relating to test-takers' performance levels. These hypotheses were constructed in light of Khalifa and Weir's (2009) model in that, under time pressure, lower-performing test-takers will focus proportionally more on lower-level processing, while higher-performing test-takers will focus proportionally more on higher-level processes. The actual metrics used to test the hypotheses in the study will be presented in the methodology section below; here we give an overview of the hypotheses, divided into three categories - overall processing, text processing and task processing. Overall processing is related to the completion of the task as a whole (text and word bank), text processing is related to the main text (i.e., the text from which the words were extracted), and task processing is related to the words in the word bank and how processing on these interacts with the main text.

Overall processing

It was hypothesised that the higher-performing test-takers would complete the tasks in less time than the lower-performing test-takers, i.e. they could do the same "cognitive work" required to resolve the gap in less time (Hypothesis M1). This is based on Khalifa and Weir's (2009) model hypothesises that word recognition is more automatised, i.e. occurs quicker, in

higher-performing test-takers, allowing more cognitive capacity to be directed towards higher-level cognitive processes needed to resolve which word fits the gap.

Text processing

It was hypothesised that higher-performing test-takers would spend proportionally more time focusing on the task's text than on other parts of the task.(Hypothesis M2). This is justified by the fact that, under time pressure, higher-performing participants would be expected to have additional cognitive resources to spend on higher-level, more global processing engendered by the sequences of words in the task's text, as opposed to local, lower-level processing, i.e. word recognition and lexical processing engendered by individual and de-contextualised words presented in the word bank.

It was also hypothesised that higher-performing test-takers would spend proportionally less time focusing on sentences containing gaps (Hypothesis M3). This hypothesis is based on the fact that the higher-performing test-takers would be expected to have additional cognitive capacity to direct towards the extraction of more global textual meaning from sentences not containing gaps, as opposed to having to prioritise lower-level processing of sentences with gaps.

Furthermore, it was hypothesised that higher-performing test-takers would spend proportionally less time focusing on the words surrounding the gaps (Hypothesis M4), which, arguably, carry the most syntactic and semantic information pertinent to the response. This is based on the fact that, similar to above, the higher-performing test-takers should be able to recognise and decode words, parse grammar and establish meaning faster than the lower-performing participants and thus have more additional cognitive capacity to direct towards the more global information not immediately surrounding the gaps that may impact on their response selection.

Task Processing

With reference to processing of the task-side, it was hypothesised that higher-performing test-takers would spend proportionally less time focusing on the word bank (Hypothesis M5). Similarly to M2, this is because, under time pressure, a smaller proportion of higher-performing test-takers' cognitive capacity would be required to decode the words in the word bank.

It was also hypothesised that higher-performing test-takers would switch their gaze less between the word bank and the text (Hypothesis M6). This is because it is expected that the additional cognitive capacity available to the higher-performing test-takers allows them to better retain the words in the word bank in their working memory. This implies less of a need for switching between the text and the word bank to "refresh" one's working memory in order to complete the gaps, whereas the opposite may be the case for the lower-performing participants, who may switch more.

Finally, it was hypothesised that the lexical frequency of a word in the word bank would have a differential processing load according to test-takers' level of performance (Hypothesis M7). Specifically, lower-performing test-takers have to attribute proportionally more cognitive resources to decoding the lower-frequency lexical items. On the other hand, it was expected that for higher-performing test-takers, who likely possess a greater lexical knowledge, a weaker relationship would be found between word frequency and fixation time as they are able to access the less frequent lexis with greater ease, thus freeing up time-limited cognitive resources for other areas of processing.

Methodology

The seven hypotheses based on Khalifa and Weir's (2009) model were investigated in a correlational study in which eye-tracking data were used to create processing measures used as independent variables and overall scores on the banked gap-fill task served as the dependent variable. This design was used to reveal the extent to which participants' engaging in higher- vs. lower-level processing is related to their performance on the banked gap-fill task type.

Participants

The participants were selected according to a stratified design with 1/3 from a pre-sessional English course, 1/3 undergraduates and 1/3 post-graduates at a British university, aiming to cover a range of English reading abilities. These three groups were sampled from to attain a sample containing variance in abilities as no external recent language proficiency data were available. Because the focus of this study was the investigation of the item type overall and not the investigation of L1 or L2 speakers specifically taking banked gap-fill items, our sample included English native speakers as well. As the native speakers were highly-educated postgraduate students, they were hoped to have comparatively high-levels of reading proficiency (at least within our sample) and provide an upper bound to the performance levels. Although this is of course not necessarily the case, their performance results (see Table 2, participants P24-P26-P27-P28) seemed to support the assumption and helped ensure variance within the present sample.

Data were collected from 28 participants who had been successfully screened for eye-tracking suitability through scanpath inspection (Holmqvist et al., 2011); eight others proved unsuitable for furnishing eye-tracking data of sufficient accuracy for the analyses. Their ages ranged between 17 and 50 years (M=31.6; SD=1.8); six were male and 22 female. They came from a variety of European and Asian backgrounds, and their first languages were: Mandarin

(12), English (4), Arabic (3), Italian (2), Sinhalese (2), German (1), Hungarian (1), Russian (1), Spanish (1) and Thai (1).

Banked gap-fill tasks

Eye-movement and reading test performance data were collected on six banked gap-fill *tasks* (i.e., full texts contacting a number of gaps) consisting of 24 *items* (i.e., individual gaps to be filled) in total, i.e. six sets of texts with each containing 3-5 gaps. In practice, we used live tasks from the Pearson Test of English Academic (PTE Academic), a computer-delivered EAP test covering the CEFR range <*A1-C2*. The tasks were developed by professional item writers and had undergone extensive review and piloting, and were thus assumed to be of good quality. They comprised a variety of topics (literature, music, philosophy, science) and represented a wide range in mean difficulty (based on the Pearson item bank difficulty values). According to Pearson's (2012) publicly available information, the targeted reading subskills of the banked gap-fill tasks used in the PTE Academic are "Identifying the topic, theme or main ideas; identifying words and phrases appropriate to the context; understanding academic vocabulary; understanding the difference between connotation and denotation; inferring the meaning of unfamiliar words; comprehending explicit and implicit information; comprehending concrete and abstract information; following a logical or chronological sequence of events" (p.25). It should be stressed, however, that the present study was not undertaken to validate PTE Academic banked gap-fill items *per se*, but rather to investigate the banked gap-fill item format in general. Figure 3 shows a freely available sample task (http://pearsonpte.com/wp-content/uploads/2014/10/Tutorial.pdf), not used in the present study.

Figure 3: Sample banked gap-fill task

In the text below some words are missing. Drag words from the box below to the appropriate place in the text. To undo an answer choice, drag the word back to the box below the text.

Master of Science in Information Technology (MSc in IT): Our programme will develop your [ ] knowledge of Computer Science and your problem-solving and [ ] skills, while enabling you to achieve the [ ] qualification for the IT professional. The programme structure is extremely [ ], enabling you to personalise your MSc through a wide range of electives.

ultimate	variable	analytical	flexible	theoretical
considerable	decisive			

The dependent variable in this study (task performance) was constructed from the participants' sum score (min=0, max=24) on the set of banked gap-fill items, whereby each gap was scored as correct=1 and incorrect=0.

Eye-tracking measures

In order to investigate test-takers' processing while completing the banked gap-fill tasks, and more specifically each of our hypotheses on the relationship between processing and performance, participants' eye traces were analysed according to seven metrics, defined in Table 1 and further explained below. It should be noted that while eye-tracking measures are

not accurate enough to map one-to-one onto components of the Khalifa and Weir (2009) model, they should allow us to make inferences about the tendencies of specific participants to engage in lower- or higher-level processing.

Table 1: Eye-tracking measures used in this study

Processing focus	**Hypothesis (H)**	**Measure (M)**	**Technical definition of the measure**
Overall processing	There is a negative relationship between overall processing time and performance (H1)	Mean time fixating on task (seconds) (M1)	The mean, across the six tasks, of the total fixation time to complete the banked gap-fill task.
Text processing	There is a positive relationship between text processing time and performance (H2)	Mean proportion of time fixating on text (M2)	The mean, across the six tasks, of the total fixation time on the text divided by the total fixation time to complete the banked gap-fill task.
	There is a negative relationship between processing time of sentences containing a gap and performance (H3)	Mean proportion of time fixating on sentences containing a gap (M3)	The mean, across the six tasks, of the total fixation time on sentences which contained a gap divided by the total fixation time on the text.
	There is a negative relationship between processing time of words surrounding the gap and performance (H4)	Mean proportion of time fixating on words surrounding gap (M4)	The mean, across the six tasks, of the total fixation time on the three words either side of the gap divided by the total fixation time on the text.
Task processing	There is a negative relationship between time spent looking at the words in the word bank and performance (H5)	Mean proportion of time fixating on word bank (M5)	The mean, across the six tasks, of the total fixation time on the word bank divided by the total fixation time to complete the banked gap-fill task.
	There is a negative relationship between frequency of text - word bank switching and performance (H6)	Mean number of visits to word bank (M6)	The mean number of transitions from the text to the word bank.
	There is a positive relationship between	Gradient of total fixation duration on	The regression slope between the total fixation duration on a

	the strength of association between word frequency and time looking at a word, and performance (H7)	word against BNC frequency (M7)	word in the word bank and the logarithm of the word's BNC frequency.

As shown in Table 1, the eye-tracking measures adopted in this study are primarily derived from the fixations made by the participants. In practice, what constituted a fixation was determined by means of a velocity-based filter (Tobii I-VT filter, *velocity threshold* – 30 degrees/second; *window length* 20ms; and *minimum fixation duration* of 60ms - see Olsen (2012) for further details and rationales for these parameters), which is considered suitable for high-speed eye-trackers (Holmqvist et al., 2011).

Six of the seven measures relate to the length of time *fixating* on particular areas in the task (M1-M5; M7). The choice for this metric was based on findings from eye-tracking research on first language reading, which has shown that at least seven variables relevant to reading processing affect fixation duration. These concern the frequency of a word (see e.g., Staub, White, Drieghe, Hollway, & Rayner (2010) who observed that "frequency and predictability influence the time the eyes spend on a word because these factors influence lexical processing itself" (p.12)), the familiarity of a word (see e.g., Williams & Morris (2004) who found that greater familiarity was associated with lower fixation durations), lexical ambiguity (see e.g., Sereno, O'Donell, & Rayner (2006) who observed longer fixation times for semantically or phonologically ambiguous words depending on contextual information), plausibility (see e.g., Rayner, Warren, Juhasz, & Liversedge (2004) who found evidence that implausible words located in a sentence attract longer fixations), contextual constraints (see e.g., Ashby, Rayner, & Clifton (2005) who noticed a tendency towards longer fixations if a word is less predictable from the preceding information), morphological effects

(see e.g., Andrews, Miller, & Rayner (2004) who found an influence of within-word effects on fixation times), and the age-of-acquisition (see e.g., Juhasz (2005) who discovered that the earlier the age-of-acquisition the lower fixation duration). Additionally, although forward and backward saccades also characterise reading, accurately utilising these for a banked-gap fill format with constant movements to the word bank is complex and challenging; hence, the absence of measures related to these metrics.

Measures M2-M4 were constructed to investigate text processing, and measures M5-M7 were used to look at task processing. For the analyses of M2-M7, a number of Areas of Interest (AOIs) were defined to capture information about eye movements relating to specific regions of the tasks/stimuli. The AOIs covered the regions of the tasks under investigation by a particular measure. So, for example, for M7 the AOIs covered individual words in the word bank and for M3 the AOIs covered three words either side of each gap. It should be noted that, although seemingly arbitrary, three words either side of the gap was felt to be a suitable cut-off. Fewer words did not sufficiently capture the grammatical information test-takers were likely to use to find the correct answer. More words, i.e. the entire clause, often took up too much text, since the texts were relatively short as a whole. Measures M2-M5 and M7 were based upon sums of fixation durations within a particular AOI. Of special mention is M6 which used the movement between the AOI containing the text and the AOI containing the whole word bank to investigate 'transitions' (Holmqvist et al., 2011), namely when the readers' gaze transfers from one AOI to another, i.e. from the text to the word bank, or *vice versa*. Such transitions can be one saccade in length, where a saccade moves directly from AOI_1 to AOI_2, or they can comprise multiple saccades, where saccades fall outside both AOIs before transition is finalised.

As mentioned already, technical definitions for each measure are provided in Table 1. However, M7 – 'Gradient of total fixation duration on word against BNC frequency' – may

require some further explanation. This measure examines the processing of the individual words in the word bank, i.e. the mean fixation time on a word, according to the natural logarithm of their British National Corpus-derived frequency (written texts, BNC*web*, CQP-Edition (BCN Web, n.d.)). For each participant, the gradient of a linear regression model describing the relationship between total fixation duration on the specific word in the word bank and its log BNC frequency of the words was calculated. To clarify, it is expected that for lower-scoring participants the gradient will tend to be negative, showing substantially more processing time on the less common words, while for the higher-scoring participants the gradient will tend towards zero, showing little processing difference between the words, with reference to their frequency. Given this, we expect a positive correlation of the gradient (of the total fixation duration on word against its BNC frequency) with performance level.

The results on these metrics constitute the independent variables in this study. The time-related metrics are based on 300 measurements of eye location per second (i.e. 300hz), with output expressed in milliseconds; however, for reasons of interpretability we will present these in seconds. Also, since to our knowledge, this is the first time most of these kinds of measures have been applied in research on language testing items, no plausible ranges for the measures are available. However, we know that by definition, the measures of frequency (M6) and time (M1) must lie between 0 and ∞, the proportions (M2, M3, M4 and M5) must lie between 0 and 1, and the regression coefficient (M7) between -∞ and ∞.

Procedure

Each participant was seated in front of a computer screen, as they would when completing the PTE Academic. Participants' eye traces were recorded with a Tobii TX300 eye-tracker – an unobtrusive, high-precision eye-tracker (300Hz sampling rate, accuracy 0.4°). The tasks were displayed in the Verdana font with a font size of 32px/24pt on a 23" monitor with an aspect

ratio of 16:9 and a resolution of 1920x1080. The line spacing was x3 normal of HTML standard to allow for vertical inaccuracy in the eye-position estimate. Each task was presented individually and no scrolling was required. Data were collected on one participant at a time.

After calibration, the participants were given a practice task in order to familiarise themselves with the banked gap-fill format. Next, they completed the six tasks, with a 15-minute time limit.

Analyses

In order to explore the relationship between the use of processes and test-takers' performance level, correlations were calculated between the eye-tracking results of each of the seven metrics (the independent variables) and participants' scores on the banked-gap fill tasks (the dependent variable). Shapiro-Wilk tests of normality indicated that all variables' distributions were not statistically significantly different from normal ($p>.05$), thus Pearson product-moment correlations were used.

Results

Table 2 shows the percentage of correctly answered banked gap-fill items by participant, and demonstrates the group's test performance range. It should be noted that percentages are provided because the eye-tracker failed to record the performance of two participants on one task; rather than remove all data from the study generated by these two participants, we felt it was more valuable to retain it and report scores as a percentage of correct responses on attempted items. As can be seen in the table, the lowest-scoring participant completed the gaps correctly in 38% of the cases, whereas three participants perfectly reconstructed all six

texts. On average, the participants answered 71% (SD=19%) of the items correctly and Cronbach's alpha was .844.

Table 2: Participants by proportion of banked gap-fill items correct

Participant	**Percentage correct (%)**	**Participant**	**Percentage correct (%)**	**Participant**	**Percentage correct (%)**
P1	38	P11	63	P21	90
P2	42	P12	63	P22	92
P3	46	P13	63	P23	92
P4	50	P14	67	P24	96
P5	54	P15	67	P25	96
P6	54	P16	75	P26	100
P7	58	P17	75	P27	100
P8	58	P18	75	P28	100
P9	58	P19	79	**Mean**	71
P10	58	P20	83	**S.D.**	19

The descriptive statistics for each of the eye-tracking measures, the results of the correlation analyses, the effect sizes, the hypotheses and an indication of support for the hypotheses are presented in Table 3. The effect sizes are interpreted according to Cohen's (1992) standard guidelines (i.e., small effect: $0.1<r<0.3$; medium effect: $0.3<r<0.5$; large effect: $r<0.5$).

Table 3: Descriptive statistics and correlations of eye-tracking measures with proportion of correct banked gap-fill items

Processing focus	Measure	M	SD	Min	Max	r	p	Effect size	Hypothesis	Hypothesis Supported?
Overall processing	Mean time on items (seconds) (M1)	67.60	25.30	30	126	-.47	.01 **	Med	There is a negative relationship between overall processing time and performance (H1)	✓
Text processing	Mean proportion of time fixating on text (M2)	.56	.04	.47	.67	-.15	.44	*NA*	There is a positive relationship between text processing time and performance (H2)	✘
	Mean proportion of time fixating on sentences containing a gap (M3)	.85	.04	.75	.93	.13	.59	*NA*	There is a negative relationship between processing time of sentences containing a gap and performance (H3)	✘
	Mean proportion of time fixating on words surrounding gap (M4)	.45	.05	.34	.54	-.59	.00 ***	Large	There is a negative relationship between processing time of words surrounding the gap and performance (H4)	✓
Task processing	Mean proportion of time fixating on word bank (M5)	.31	.04	.23	.40	.07	.72	*NA*	There is a negative relationship between time spent looking at the words in the word bank and performance (H5)	✘
	Mean number of visits to word bank (M6)	17.80	6.19	5.83	33.83	-.54	.00 **	Large	There is a negative relationship between frequency of text - word bank switching and performance (H6)	✓
	Gradient for total fixation duration on word against BNC frequency (M7)	-.47	.52	-2.23	.14	.42	.03 *	Med	There is a positive relationship between the strength of association between word frequency and time looking at a word, and performance (H7)	✓

* significant at p≤.05, ** significant at p≤.01, and *** significant at p≤.001

It should be noted that when we discuss 'higher' and 'lower' performing test-takers we are referring to the general tendencies of those at either end of the score distribution on the banked gap-fill tasks rather than explicitly comparing two groups. The significant negative association between banked-gap fill scores and the average time spent completing each task (r=-.47, p=.01*), with medium effect size, is in line with our expectation that higher-performing test-takers would complete the task more quickly than lower-performing test-takers (Hypothesis M1).

Regarding the text processing, evidence was not found for the tendency of higher-scoring test-takers to spend proportionally more time processing the text (compared to processing other parts of the task) than lower-scoring test-takers (r=-.15, p=.44). Also contrary to what was anticipated, the mean proportion of time fixating on sentences that contain a gap did not correlate significantly with banked gap-fill performance (r=.13, p=.56). Thus our hypotheses that lower-performing test-takers would spend proportionally more time on the texts (Hypothesis M2) and on the sentences with the gaps (Hypothesis M3) were not supported. Nevertheless, those test-takers who scored lower did spend proportionally more time focussing on the three words around the gaps, as shown by the strong significant negative correlation between 'mean proportion of time fixating on words surrounding gap' and test-taker performance (r=-.59, p=.00***) – a large effect size. Thus, while slightly broader-scope processing at sentence level required rather similar proportional processing times for both higher and lower performers, lower-performing test-takers did spend proportionally more time reading locally around the gaps, assumedly parsing the immediate grammatical and lexical context (Hypothesis M4).

In terms of task processing, the average proportion of time test-takers fixated on the word bank of each task did not significantly correlate with their performance results (r=.07, p=.72) (Hypothesis M5). However, the frequency with which they shifted their attention from

the text to the word bank did correlate negatively with their scores (r=-.54, p=.00**) – with a large effect size. More specifically, the better performers made fewer visits to the word bank than did those with lower scores on the banked gap-fill tasks, indicating more fragmented processing by lower-scoring test-takers, possibly a result of reduced or lack of processing capacity to effectively store the words from the word bank in their working memory (Hypothesis M6). Furthermore, lower-scoring test-takers spent more time on individual words in the word bank according to the words' BNC-derived frequency (r=.42, p=.03*), with a medium effect size, thus probably dedicating more processing time to these and leaving proportionally less capacity available for other types of processing (Hypothesis M7).

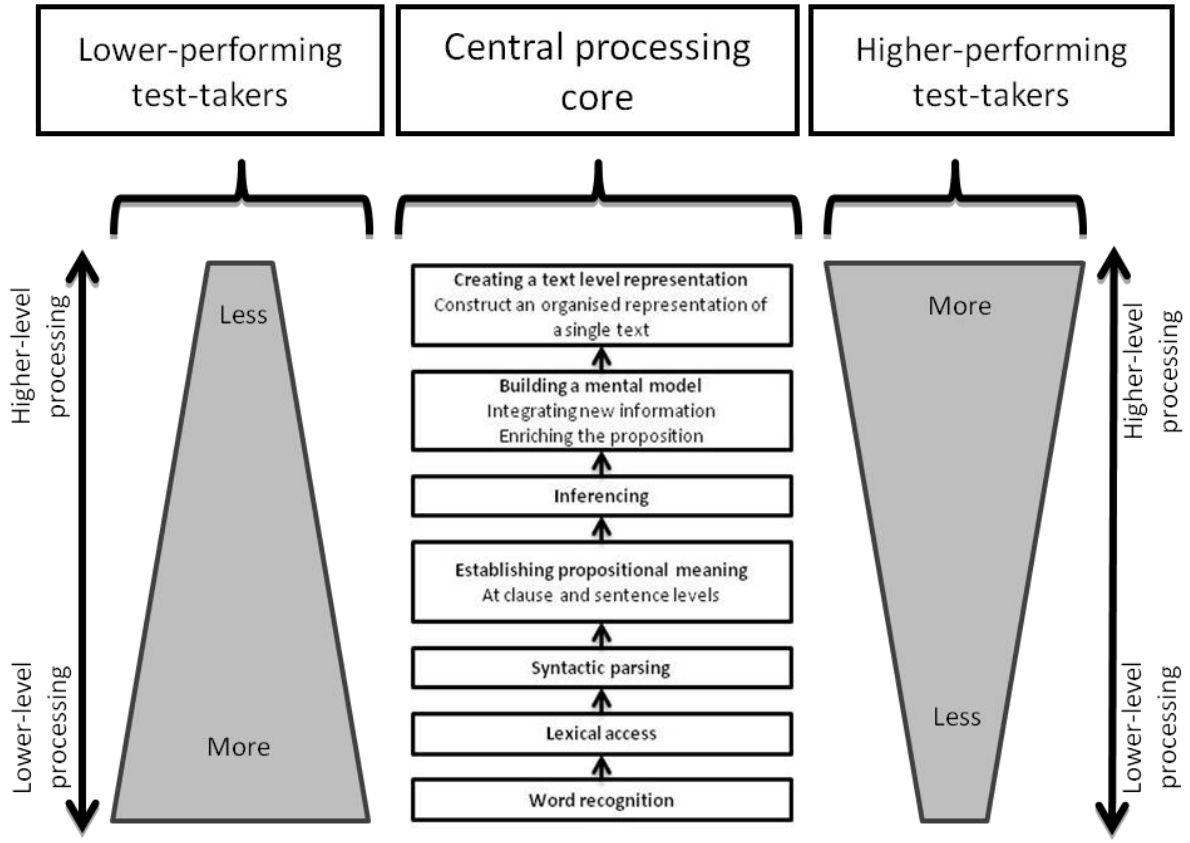
Discussion

The findings presented above suggest that the application of eye-tracking can indeed give some insight into cognitive processing while completing banked gap-fill tasks designed to assess reading ability and into differences in processing profiles between higher- and lower-scoring test-takers. This strengthens the evidence for the usefulness of this method, as reported in Bax and Weir (2012), Bax (2013), and Brunfaut & McCray (2015), to investigate test-takers' reading processes while completing items of various task types (e.g. multiple-choice, matching, sentence completion). In particular, the present study suggests differences in cognitive processing profiles across reading performance levels, with less successful test-takers demonstrating more evidence of the usage of lower-level processing in their responses to the banked gap-fill items.

Mapped onto the central processing core of the Khalifa and Weir (2009) framework this could be visualised as shown in Figure 4. It should be noted, however, that the Figure does not aim to provide a detailed description of the cognitive processing profiles across higher- and lower-performing test-takers. Rather, it attempts to communicate the study's key

finding of a probable greater emphasis on lower-level cognitive processing by lower-performing test-takers. Namely, the widths of the rhombuses for lower- and higher-performing test-takers are representative of the relative proportion of processing time at the different levels of the Khalifa and Weir (2009) central processing core. Also, while lower-performing test-takers did exhibit more of a tendency towards more local, lower-level processing, it does not mean that they did not engage in more global, higher-level processing at all, and *vice versa*. Indeed, the double-sided arrows in Figure 4 are indicative that both top-down and bottom-up processing is occurring. Overall, though, the evidence suggests that test scores for lower-performing test-takers on banked gap-fill tasks may primarily show their competence in the lower-level processes of reading rather than reading as a whole. It is important to state, however, that this differential processing profile may not be a problem if the scores on a test are claiming to represent the ability of a test-taker to engage in higher-level, more global reading processing; in fact, the results show that this item type is, to some extent, measuring this aspect of performance.

Figure 4: Visual representation of proportional processing time between higher- and lower-performing test-takers completing banked gap-fill items (adapted from Khalifa & Weir, 2009, p.43)



The position that lower-performing test-takers engaged proportionally more in lower-level cognitive processing has been arrived at following three key findings. Firstly, it is based on the fact that lower-scoring test-takers focused proportionally more on the three words surrounding the gaps than did higher-scoring test-takers (M4). The increased attention to the words immediately surrounding the gaps is indicative of more localised reading, which implies more lower-level cognitive processing. The second source of evidence relates to the finding that lower-scoring test-takers made more visits to the word bank than did higher-scoring test-takers (M6). More frequent transitions from the text to the word bank indicate more disruption to the 'linear' reading process. As bottom-up cognitive processes in reading are hypothesised to build upon one another (see Khalifa & Weir's (2009) central processing

core in Figures 1 & 4), and words are being parsed into proposition units, and proposition units are being combined into a mental model of the text, etc., disruption to this process indicates an increased likelihood of more of the processing occurring at the lower-levels. Thirdly, lower-scoring test-takers spent proportionally more time on less frequent word options in the bank (M7), thereby dedicating more resources to lower-level processes such as recognising and accessing the words listed in the word bank, leaving fewer resources available for other forms of processing.

These findings provide some support for Brown's (2004) hypothesis that the way in which gap-filling items are processed is related to reading performance, albeit only based on reading performance on banked gap-fill items. Additionally, this study's findings are in line with those of Yamashita (2003), whose study – although limited in scope – found support for Brown's hypothesis in the context of open gap-fill items. Our findings seem to provide additional evidence for Brown's hypothesis, with reference to banked gap-fill items and based on a larger dataset investigated with eye-tracking methodology.

It should be noted, however, that some of the eye-tracking metrics used did not point to significantly different cognitive processing profiles depending on test-takers' reading performance. No significant difference was revealed in the cognitive processing profile across higher- and lower-scoring test-takers for sentences containing a gap (M3), which was contrary to our expectation of processing differences according to test-taker performance at the sentential level. This finding also contrasted with the differences in cognitive processing profiles for even more local reading, namely of the three words surrounding the gaps. However, a likely explanation for this is that the banked gap-fill texts only contained one sentence that did not have a gap (and two had no such sentence), thus making the measurement of this metric more prone to random error. Furthermore, although no systematic processing profile differences were established between higher- and lower-performing test-

takers, in terms of the proportion of time they spent focusing on the text (M2) and the word bank (M5), it should be noted that there was a lot of individual variance in these measures. For example, for the word bank fixation time, some participants spent approximately 20% of their time looking at the word bank, but others spent approximately 40% of their total task completion time focusing on it. This seems to suggest that the amount of time spent processing the text itself or vocabulary listed in the word bank does vary significantly from individual to individual but is not related to performance. Potentially, factors such as concentration level, test-taking habits or strategies (e.g., targeted search strategy for a particular word in the bank versus each time running through all words in the word bank), or working memory (ability to keep words in the bank in mind at all times and thus less time needed to identify the word one is looking for when visiting the word bank) may play a role in this.

Apart from implications for what aspects of reading processing banked gap-fill items tend to measure depending on test-takers' ability, the present study's findings also suggest the necessity for careful task design in banked gap-fill items to avoid too much shifting from assessing a full range of lower- and higher-level processing, underlying proficient reading ability, to only lower-level processing or even only testing lexico-grammar in context. In particular, the analysis of the relationship between the word bank options and their BNC-derived frequency, which showed more processing occurring on the less frequent words, suggests that the use of more complex lexis modifies the cognitive processing profile of the test for lower-performing test-takers towards vocabulary, i.e. word recognition and lexical access. This finding shows that the frequency of the word deleted from the text or selected as a distractor can influence the amount of processing it receives, and that this differs according to test-taker ability. Thus, if used for testing reading beyond low-level lexical processes, it might be preferable to carefully consider the selection of less complex vocabulary when

forming gaps and also opt for less complex vocabulary to serve as distractors while keeping the difficulty of the text constant. In this manner, the items may be less dependent on knowledge of vocabulary and potentially more able to elicit a larger range of the cognitive processes involved in reading for all levels of test-takers. In addition, the finding that lower-scoring test-takers conducted more local parsing around the gaps underlines the importance of supplying enough grammatically acceptable distractors for each gap. A lack of these is likely to shift the processing more to syntactic parsing than to the full range of cognitive processes involved in reading, in particular for lower-proficiency test-takers. Thus, careful task design is needed to ensure valid use of the banked gap-fill format to assess reading ability in general and of test-takers of various levels of ability. In this respect, overtly manipulating gaps and distractors forms an interesting avenue for research on task design effects on the construct being tested.

Conclusion

This study has illuminated the cognitive processing conducted by test-takers while completing banked gap-fill tasks used to assess reading. Importantly, it has shown a number of salient processing differences depending on test-takers' performance level, which, as discussed above, have implications for the design of banked gap-fill tasks to warrant valid reading assessment. In particular, the data indicated that a major difference in the processing between lower- and higher-scoring test-takers is related to the local context of the gap and to the complexity of words in the word bank, and that lower-scoring test-takers thereby rely more on lower-level cognitive processing in their responses to the banked gap-fill items.

One limitation of the study concerns its size. Although based on a much larger set of tasks (i.e., six tasks in this study) than past process-oriented research (typically one task per study), and constrained by the fact that eye-tracking research typically includes a limited

number of participants due to its resource-heavy nature, a larger sample size would enable more complex analyses (i.e., multivariate statistical analyses such as multiple regression, mixed effects models, factor analysis or SEM). Additionally, to explore test-takers' overall text- and task-processing while completing banked gap-fill tasks, the study made use of eye-tracking, which proved to be a useful method. However, in order to gain an even more detailed understanding of the potential of this task type for testing reading – especially insights into the exact cognitive processes being employed during banked gap-fill completion, it would be meaningful to follow-up this study with research that employs a mixed-methods design. This could for example involve the use of much more sophisticated eye-tracking metrics that provide evidence for, in particular, very specific lower-level cognitive processes, and complements this with verbal report data such as stimulated recalls to provide evidence especially on very specific higher-level processes such as inferencing (which cannot be observed with eye trace data) and on the interaction between processes. A promising example of such a mixed-methods methodology is provided in Brunfaut & McCray (2015).

Acknowledgments

This work was supported by the Economic and Social Research Council Advanced Quantitative Methods Postgraduate Studentship (grant number M84810K). We wish to thank Pearson Language Testing for providing us with test tasks. We are also grateful to Em. Prof. J. Charles Alderson for his continued support during the research reported in this article.

References

Alderson, J. C. (1980). Native and non-native speaker performance on cloze tests. *Language Learning*, *30*(1), 219–223.

Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

Alderson, J. C., & Csereszнyés, M. (2003). *Into Europe: Reading and use of English*. Budapest: Teleki László Foundation.

Anderson, N. J., Bachman, L., Perkins, K., & Cohen, A. (1991). An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. *Language Testing*, *8*(1), 41–66.

Andrews, S., Miller, B., & Rayner, K. (2004). Eye movements and morphological segmentation of compound words: There is a mouse in mousetrap. *European Journal of Cognitive Psychology*, *16*, 258–311.

British Council (n.d.) Aptis - Flexible English Test. Last accessed July 07, 2016, from https://www.britishcouncil.org/exam/aptis

Ashby, K., Rayner, J., & Clifton, C. (2005). Eye movements of highly skilled and average readers: Differential effects of frequency and predictability. *Quarterly Journal of Experimental Psychology*, *58*(6), 1065–1086.

Brunfaut, T., & McCray, G. (2015). *Looking into test-takers' cognitive processes whilst completing reading tasks: a mixed-method eye-tracking and stimulated recall study*. (ARAGs Research Reports Online; Vol. AR/2015/001). London: The British Council. https://www.britishcouncil.org/sites/default/files/brunfaut_and_mccray_report_final_0.pdf

Bachman, L. F. (1985). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly*, *19*(3), 553–556.

Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye-tracking. *Language Testing*, *30*(4), 441–465.

Bax, S., & Weir, C. J. (2012). Investigating learners' cognitive processes during a computer-based CAE reading text. *Research Notes*, *47*, 3–14.

Bowles, M. E. (2010). *The think-aloud controversy in second language research*. New York: Routledge.

BNCweb. (n.d.). Last retrieved July 07, 2016, from http://corpora.lancs.ac.uk/BNCweb/

Brown, J. D. (1983). A closer look at cloze: Validity and reliability. In J. W. Oller (Ed.), *Issues in language testing research* (pp. 237–250). Rowley, MA: Newbury House.

Brown, J. D. (2004). Twenty-five years of cloze testing research: So what? In G. Poedjosoedarmo (Ed.), *Teaching and assessing language proficiency*. Signapore: SEAMEO Regional Language Centre.

Chapelle, C. A., & Abraham, R. G. (1990). Cloze method: What difference does it make? *Language Testing, 7*(2), 121–146.

Chihara, T., Oller, J., Weaver, K., & Chavez-Oller, M. A. (1977). Are cloze items sensitive to constraints across sentences? *Language Learning, 17(1)*, 63–70.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159.

Cziko, G. A. (1978). Differences in first and second language reading: The use of syntactic, semantic and discourse constraints. *Canadian Modern Language Review, 34*, 473–489.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing*. Cambridge: Cambridge University Press.

Engelhard, G. (2001). Historical views of the influences of measurement and reading theories on the assessment of reading. *Journal of Applied Measurement, 2*(1), 1–28.

Gamarra, A., & Jonz, J. (1987). Cloze procedures and the sequence of text. In R. S. Readence & R. S. Baldwin (Eds.), *Research in literacy: Merging perspectives* (pp. 17–24). Rochester, NY: National Reading Conference.

Gao, X., & Gu, X. (2008). An introspective study on test-taking process for banked cloze. *CELEA Journal, 31*(4), 3–16.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. New York: Cambridge University Press.

Holmqvist, K., Nyström, M., Anderson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). *Eye tracking*. Oxford: Oxford University Press.

Ilc, G., & Stopar, A. (2015). Validating the Slovenian national alignment to CEFR: The case of the B2 reading comprehension examination in English. *Language Testing, 32*(4), 443–462.

Jonz, J. (1976). Improving on the basic egg: the m-c cloze. *Language Learning, 26*(2), 255-265.

Jonz, J. (1987). Textual cohesion and second language comprehension. *Language Learning, 37*(3), 409–438.

Jonz, J. (1990). Another turn in the conversation: What does cloze measure? *TESOL Quarterly, 24*(1), 61–83.

Juhasz, B. (2005). Age-of-acquisition effects in word and picture identification. *Psychological Bulletin, 131*(5), 684–712.

Khalifa, H., & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading*. Cambridge: Cambridge University Press.

Kibby, M. W. (1980). Intersentential processes in reading comprehension. *Journal of Reading Behaviour, 12*(4), 299–312.

Klein-Braley, C. (1983). A cloze is a cloze is a question. In J. W. Oller (Ed.), *Issues in language testing research* (pp. 218-228). Rochester, NY: Newbury House.

Leys, M., Fielding, L., Herman, P., & Pearson, P. D. (1983). Does cloze measure intersentence comprehension? A modified replication of Shanahan, Kamil and Tobin. In J. A. Niles & L. A. Harris (Eds.), *Searches for meaning in reading/language processing and instruction* (pp. 111–114). Rochester, NY: National Reading Conference.

Markham, P. L. (1985). Rational deletion cloze and global comprehension in German. *Language Learning, 35*(3), 423–430.

McKenna, M. C., & Layton, K. (1990). Concurrent validity of cloze as a measure of intersentential comprehension. *Journal of Educational Psychology, 82*(2), 372–377.

Miller, G. R., & Coleman, E. B. (1967). A set of thirty-six passages calibrated for complexity. *Journal of Verbal Learning and Verbal Behavior, 6*(6), 851–854.

Oller, J. W. (1975). Cloze, discourse and approximations to English. In M. K. Burt & H. Dulay (Eds.), *New directions in second language teaching and bilingual education* (pp. 345–355). Washington, DC: TESOL.

Oller, J. W., & Jonz, J. (1994a). Why cloze Procedure? In J. W. Oller. & J. Jonz (Eds.), *Cloze and coherence* (pp. 1-20). Cranbury, NJ: Bucknell University Press.

Oller, J. W., & Jonz, J. (1994b). A comprehensive theory of coherence and cloze research. In J. W. Oller. & J. Jonz (Eds.), *Cloze and coherence* (pp. 48-80). Cranbury, NJ: Bucknell University Press.

Oller, J. W., & Jonz, J. (1994c). A review of theories of coherence. In J. W. Oller. & J. Jonz (Eds.), *Cloze and coherence* (pp. 21-47). Cranbury, NJ: Bucknell University Press.

Olsen, A. (2012). *The Tobii I-VT fixation filter*. Retrieved from http://www.acuity-ets.com/downloads/Tobii%20I-VT%20Fixation%20Filter.pdf

Owen, N. (2015, November). *Using video-stimulated recall interviews to provide cognitive validity evidence in L2 reading test*. Paper presented at LTF, Oxford, UK.

Pearson (2012). *PTE Academic score guide (version 4)*. Retrieved from http://pearsonpte.com/wp-content/uploads/2015/11/PTEA_Score_Guide_05Nov15.pdf

Pearson (n.d.) PTE Academic - The English test that takes you places. Last accessed July 07, 2016, from http://pearsonpte.com/

Pollitt, A., & Taylor, L. (2006). Cognitive psychology and reading assessment. In M. Sainsbury, C. Harrison, & A. Watts (Eds.), *Assessing reading from theories to classrooms* (pp. 38–49). Slough: NFER.

Porter, D. (1978). Cloze procedure and equivalence. *Language Learning, 28*(2), 333–341.

Porter, D. (1983). The effect of quantity of context on the ability to make linguistic predictions: A flaw in the measure of general proficiency. In A. Hughes & D. Porter (Eds.), *Current developments in language testing* (pp. 63–74). London: Academic Press.

Perfetti, C. A. (1985). *Reading ability*. Oxford: Oxford University Press.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 372–422.

Rayner, K., Juhasz, B., & Pollatsek, A. (2007). Eye movements during reading. In M. Snowling & C. Hulme (Eds.), *The science of reading* (pp. 79–97). Malden, MA: Wiley-Blackwell.

Rayner, K., Pollatsek, A., Ashby, J., & Clifton, C. (2012). *Psychology of reading* (2nd ed.). New York: Psychology Press.

Rayner, K., Warren, T., Juhasz, B. J., & Liversedge, S. P. (2004). The effect of plausibility on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(6), 1290–1301.

Sasaki, M. (2000). Effects of cultural schemata on students' test-taking processes for cloze tests: A multiple data source approach. *Language Testing, 17*(1), 229–255.

Sereno, S. C., O'Donnell, P. J., & Rayner, K. (2006). Eye movements and lexical ambiguity resolution: Investigating the subordinate bias effect. *Journal of Experimental Psychology: Human Perception and Performance, 32*(2), 335–350.

Shanahan, T., & Kamil, M. L. (1983). A further comparison of sensitivity of cloze and recall to passage organization. In J. A. Niles & L. A. Harris (Eds.), *Searches for meaning in reading/language processing and instruction* (pp. 123–128). Rochester, NY: National Reading Conference.

Soudek, M., & Soudek, L. I. (1983). Cloze after thirty years: new uses in language teaching. *ELT Journal, 37*(4), 335–340.

Staub, A., White, S. J., Drieghe, D., Hollway, E. C., Rayner, K. (2010). Distributional effects of word frequency on eye fixation durations. *Journal of Experimental Psychology: Human Perception and Performance, 36*(5), 1280–1293.

Taylor, W. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly, 30*, 415–453.

Taylor, W. (1957). "Cloze" readability scores as indices of individual differences in comprehension and aptitude. *Journal of Applied Psychology, 41*(1), 19–26.

Urquhart, A. H., & Weir, C. J. (1998). *Reading in a second language: Process, product and practice*. London/New York: Longman.

Wagner, R. & Stanovich, K. (1996). Expertise in reading. In K.A. Ericsson (Ed.), *The road to excellence: The acquisition of expert performance in the Arts and Sciences, Sports and Games* (pp. 189-225). Mahwah, NJ: Erlbaum.

Weir, C. J. (2005). *Language testing and validation*. Basingstoke: Palgrave Macmillan.

Weir, C., Hawkey, R., Green, A., & Devi, S. (2012). The cognitive processes underlying the academic reading construct as measured by IELTS. *IELTS Research Reports, 9*, 157–189.

Williams, R. S., & Morris, R. K. (2004). Eye movements, word familiarity, and vocabulary acquisition. *European Journal of Cognitive Psychology*, *16*(1-2), 312–339.

Yamashita, J. (2003). Processes of taking a gap-filling test: Comparison of skilled and less skilled EFL readers. *Language Testing*, *20*(3), 267–293.