Item writing skills and their development: Insights from an induction item writer training course

Olena Rossi

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

Department of Linguistics and English Language
Lancaster University

September 2020

Declaration

I declare that this thesis is the result of my own research except as cited in references. The thesis has not been submitted in candidature of any other degree.

Excerpts of this thesis have been published or accepted for publication in the following academic articles:

Rossi, O., & Brunfaut, T. (2019). Test item writers. In J. I. Liontas (Ed.), *The TESOL Encyclopaedia of English Language Teaching* (pp.1-7). John Wiley & Sons. DOI: 10.1002/9781118784235. eelt0981

Rossi, O., & Brunfaut, T. (2021). Text authenticity in listening assessment: Can item writers be trained to produce authentic-sounding texts? *Language Assessment Quarterly*. DOI: 10.1080/15434303.2021.1895162

Signature:

Name: OLENA ROSSI

Date: 26.02.2021

To my dear husband Peter Antonio Rossi

Abstract

Item quality makes a significant contribution to test validity, thus rendering the work of item writers critically important for assessment. However, little empirical research has so far been done into item writing, including item-writing training. This thesis therefore aimed to investigate an online induction item-writing training course in order to gain insights into the nature of item-writing skills and their development.

This research project, which is based on an existing item-writing training course, adopted a mixed-methods approach consisting of a pretest-posttest quasi-experimental study and a course feedback study. To investigate how the quality of items produced by participants changed from before to after the training (RQ1), 25 trainees produced grammar MC items, writing prompts, and listening tasks for the pre- and post-training assignments. The quality of items was evaluated by expert item reviewers against an evaluation scale; the evaluations were then analysed statistically to identify changes in item quality and individual item-writer variation. To investigate how the participants' item-writing skills developed through the training (RQ2), interviews were conducted with willing participants upon completion of each assignment and analysed using the Grounded Theory approach. Finally, to identify what role the training played in the participants' item-writing skill development (RQ3), participants' reactions to the course were collected via four feedback questionnaires administered throughout the course and analysed using quantitative and qualitative methods.

It was found that the total post-training scores for the grammar items and for the listening tasks were statistically significantly higher compared to the pre-training ones, largely due to an improvement in quality on objectively-scored criteria. Three main participant profiles were identified: (a) those whose item quality was low prior to the training but who produced better quality items following it; (b) those who produced good quality items before the training and whose post-training items were of even better quality; (c) those whose pre-training items were of reasonably good quality but whose post-training items scored one or several points lower. The analysis of interview transcripts showed that awareness of objective requirements and the ability to use item-writing tools were generally sufficient in complying with these requirements. For subjective requirements, however, the analysis revealed different approaches to item writing by participants in different profile groups. The course features that the participants reported as most useful for developing their item-writing skills included: input in language assessment principles, balance of theory and practice, variety of activities, extensive item-writing practice, and detailed feedback on items.

The findings for the three research questions were then triangulated to provide rich insights into the nature of item-writing skills and their development. The findings were interpreted with reference to two learning theories – cognitive ACT-R theory (Anderson, 1993) and social Communities of Practice (CoP) theory (Love & Wenger, 1991; Wenger, 1998). It was found that item-writing skills are item-type and proficiency-level specific and consist of multiple components acquired at different rates. It was further found that, while item-writing skill acquisition follows the 'typical' process of complex cognitive skills' acquisition as described in ACT-R (Anderson, 1993), the trajectories of acquisition for individual trainees might vary, with three main trainee profiles described. Finally, this study's findings confirmed that item writers are a CoP, and elements of legitimate peripheral participation (Love & Wenger, 1991) in an item-writers' community make item-writing training more effective.

This study contributes to understanding the nature of item-writing skills and their development through induction training. The study also advances the methodology of research into item-writing training effects. From a practical perspective, this study provides a range of recommendations concerning operational item writing, item-writing training, and item-writer recruitment.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Prof. Tineke Brunfaut, for the unwavering support she provided throughout my time as a doctoral student. She was the best supervisor any student can wish for. If, as I hope, my research and academic writing skills have developed throughout these years, this is solely due to her guidance.

I would also like to thank Prof. Luke Harding and Dr. John Pill for their valuable advice as departmental panellists during my PhD journey: Prof. Harding's insightful comments during the confirmation panel, and Dr. Pill's valuable suggestions during the post-confirmation panel helped improve this research project. I also wish to express my appreciation to the members of the Language Testing Research Group (LTRG) at Lancaster University, who attended my presentations on different aspects of this research project, for their useful discussions and suggestions. Thanks also go to Dr. Michael Ratajczak for his help with statistics.

I am grateful to the British Council Assessment Research Group for granting me the Assessment Research Award in 2018. Without this financial support, it would have been difficult to carry out the study.

My sincere thanks go to all participants of this study. This thesis would not have been possible without the item-writing trainees and expert judges who contributed their time and effort to make this study possible. I would also like to thank my former colleagues and managers Sheryl Cooke, Philip Horne, Colin Barnett, and Judith Fairbairn for their encouragement and useful discussions.

Above all, my most heartfelt and everlasting thanks go to my husband Peter. Without him, there would be no study and no thesis: his encouragement set me on this journey, while his dedication and personal sacrifices allowed me to complete it.

Contents

List of tables

List of figures

List of abbreviations

ACT-R	Adaptive Control of Thought – Rationale
CEFR	Common European Framework of Reference
CoP	Community of Practice
DCoP	Distributed Community of Practice
FQ1	Feedback Questionnaire 1
FQ2	Feedback Questionnaire 2
FQ3	Feedback Questionnaire 3
FFQ	Final Feedback Questionnaire
GR	Gain Ratio
MC	Multiple Choice
RQ	Research Question
VLC	Virtual Learning Community

List of appendices

Chapter 1 Introduction

1.1 Rationale for the study

Item writers are those people who produce test items, normally according to a set of specifications, to make up a test. As item writers effectively fill the test with content, their work is absolutely vital to testing. Indeed, Bachman and Palmer (2012) stated that "task writers are key personnel in the assessment development process" (p.417). Moreover, it has been repeatedly emphasized in the language testing and educational measurement literature (Bachman, 1990; Bachman & Palmer, 2012; Haladyna & Rodriguez, 2013; Lane et al., 2015; Messick, 1996; Weir, 2005) that the quality of test items is of crucial importance for test validity and that item validation should be an integral part of the test validation process.

Although item quality is recognised as vital for test validity, little is known about the people who produce items, and the work they do. Bachman and Palmer (2012) devoted only 15 lines to item writers in their seminal, 500-page book *Language assessment in practice*. This is typical in language assessment – item writers are often viewed as quasi-professionals and their work receives much less attention than that of test designers, examiners, raters, or test data analysts (Shin, 2012).

It is also unclear from the literature what the nature of item-writing expertise is, how one becomes an expert at writing items, and what role training plays. For example, does one become an expert item writer after an X number of years writing items? And if so, for how many years should one be producing test items to be regarded as an expert? Or does it depend on the work outcomes, that is how many items a person has written and how many of those items have been accepted for live testing? Scholarly sources that touch upon item-writing expertise do not agree, with some of them prioritising years of professional practice but giving different answers as to how many years are required to be considered an item-writing expert (Fulkerson et al., 2010; Johnson et al., 2017), while others see the number of successful commissions as more important (e.g., Green & Hawkey, 2011; Salisbury, 2005).

Very few studies exist that have investigated the process of writing test items, both within the language testing field (Green & Hawkey, 2012; Kim et al., 2010; Salisbury, 2005) and within educational measurement in general (Fulkerson et al., 2009; Johnson et al., 2017). The studies that exist have not resulted in a comprehensive account of item-writing skill development –

they were mostly framed as experience sharing and looked at the item-writing process of either experienced (Johnson et al., 2017) or inexperienced (Kim et al., 2010) item writers, or compared the item-writing processes of experienced versus inexperienced item writers (Green & Hawkey, 2012; Salisbury, 2005), at one point in time.

It is often the case in language testing that item writers do not receive any formal training but have to learn to write items by writing them (Alderson, 2010). For example, language teachers in all types of educational establishments – from schools to universities – are expected to write a range of language tests/assessments with very little or no training and limited item-writing skills. Although many exam boards do provide in-house training to their item writers, who are usually freelancers, little information is available in the public domain on how this training is organised. When available (e.g., Ingham, 2008; de Jong, 2007), critical reflection and evaluation of training effectiveness is lacking. This situation is particularly surprising given that it has been recognised in the educational measurement literature that training item writers "constitutes evidence for item validation" (Haladyna & Rodriguez, 2013, p.22).

Scholarly sources provide little to no advice on how to organise item-writing training or how to measure its effectiveness. Although language testing textbooks give some practical recommendations on item writing in order to produce good-quality items (e.g., Alderson, Clapham & Wall, 1995; Brown, 2010; Heaton, 1990; Hughes, 2003), they contain no guidance on how people can be trained in item writing. Literature in the field of educational measurement provides comparatively more information on this topic (Haladyna & Rodriguez, 2013; Welch, 2006), but the training recommendations are generally based on the authors' practical experience rather than on empirical research into item-writing skills and how to develop these effectively through training. To the best of my knowledge, no empirical proof of the effectiveness of item-writing training has been presented in the language testing literature, while the few recent studies coming from other fields (e.g., Abdulghani et al., 2015; Dellinges & Curtis, 2017; Gupta et al., 2020; Hamamoto Filho & Bicudo, 2020) produced unconvincing results due to multiple methodological flaws in the studies concerned (see Section 2.2.7.1). The latter also points to the fact that the methodology for researching item-writing training is still in its infancy.

This lack of research into item-writing skill development does not reflect the state of research into skills' development more generally. Multiple learning theories have been put forward representing different approaches to investigating skill development. Some of these theories consider skill acquisition as an individual cognitive process (e.g. Adaptive Control of Thought theory, Anderson, 1993) while others look at skill development as a social endeavour (e.g., Communities of Practice theory, Lave & Wenger, 1991). The process of skill acquisition has been empirically investigated, for example, for motor skills (Fitts, 1964), X-ray picture diagnosing (Lesgold et al., 1988), text editing (Singley & Anderson, 1985), computer programming (Anderson, 1987), to name just a few. In applied linguistics, multiple models of child language acquisition (see e.g., Ambridge & Lieven, 2011) and second language acquisition (see e.g., Vanpatten & Williams, 2015) exist. It seems surprising, then, that the process of item-writing skill development has been largely neglected. Besides offering interesting theoretical insights into how an item writer develops from a novice to an expert, such insights would arguably be of great practical benefit to inform item-writing training and to set realistic expectations for how quickly one can develop into a skilled item writer.

These gaps in research into item-writing skill development were my main motivation to conduct the present study. However, this motivation was also reinforced through years of personal experience in this area. My background is in teaching English as a foreign language, having 17 years of experience teaching learners of different ages and proficiency levels in four countries. While working as a teacher, and then teacher trainer, my interest gradually shifted in the direction of language testing: I prepared students for international English exams and qualified as an examiner for several international exams. I also had to create tests for my students without having a clear idea of how to do that well. This left me wondering how item writing (I did not know the term then) was done professionally. Then, in 2012, I was given the opportunity to be trained as an item writer; my practical involvement with item writing continues until now – I am regularly commissioned for a range of items and have also become an item reviewer and item-writing trainer. Although I received some practical training in item writing, I felt that my knowledge of theory and principles of language assessment was still lacking, which I thought was negatively affecting my item-writing practice. For example, I often wondered why I was asked to create items in a particular way – the explanation that this was required in the specifications did not satisfy me. Therefore, I studied for an MA in Language Testing at Lancaster University and then started a PhD at the same university, taking item writing as the topic for my doctoral study.

1.2 Background to the study

This research project was carried out during the second and third run (Cohorts 2 and 3) of an online item-writing training course provided to employees of the British Council China. The aim of the course was to train existing employees who would then possibly contribute to various assessment projects in which the organisation was involved in the East-Asia region. The training course was not created for the purpose of this research project: the intention to investigate the development of item-writing skills through a course came somewhat later, with Cohort 1 having completed their training before the research project began.

The sub-sections that follow describe the general basis on which the course was developed: the view on item-writing skill development that informed the creation of this course, the theoretical framework used to produce item specifications for the training, and the training principles that informed the course design. The final sub-section provides some general information about the course participants and their motivation for taking the course (for more detailed information see Methodology Chapter, Section 3.2).

1.2.1 The view on item-writing skill development that informed course design

The APA Dictionary of Psychology defines 'skill' as "an ability or proficiency acquired through training and practice" (American Psychological Association, n.d.). Following Davies et al.'s (1999) definition of item writing as "the stage of test development in which test items are produced, according to a set of test specifications" (p.99), I define 'item-writing skill' in this thesis as the ability to produce test items according to a set of test specifications, the ability which is acquired though relevant training and/or item-writing practice. Although many practicing item writers have never received formal training (see, e.g., Alderson, 2010; an overview of the literature on item-writing training is presented in Section 2.2.6 of this thesis), the item-writing course researched in this study was designed with the belief that prospective item writers need to be trained before they can be commissioned to produce test items. Therefore, the overarching aim of this training course was to develop trainees' ability to produce test items according to specifications.

The scope of the training course was defined based on the notion of language assessment literacy (LAL). Fulcher (2012) provided a detailed working definition of LAL as

> The knowledge, skills and abilities required to design, develop, maintain or evaluate, large-scale standardized and/or classroom based tests, familiarity with test processes, and awareness of principles and concepts that guide and underpin practice, including ethics and codes of practice (p.125)

This definition emphasizes multidimensionality of LAL, which includes both the theoretical knowledge of language testing principles and the practical ability in test development. The item-writing training course researched in this study was thus designed to develop both aspects of LAL in the trainees. The main course objectives were defined as: (1) equipping trainees with knowledge of language testing principles relevant to the job of item writing, and (2) developing the trainees' practical ability to produce language test items against existing specifications. The programme was not aimed at writing items for a specific language test but was developed with general, large-scale English language proficiency testing in mind. To this end, the programme included training in producing a wide range of item types, both selected- and constructed-response, to test grammar, vocabulary, and the four language skills (reading, writing, listening, and speaking).

Pill and Harding (2013), drawing on literacy models in the fields of mathematical and scientific education, described five stages of LAL development, from illiteracy through functional literacy to multidimensional literacy (p.383). Taylor (2013), capitalising on the multidimensional developmental view of LAL proposed by Fulcher (2012) and Pill and Harding (2013), speculatively described eight-dimensional LAL profiles for several stakeholder groups, including one for test writers (i.e. test developers) (*Figure 1-1*). Most recently, these dimensions were empirically tested by Kremmel and Harding (2020), who arrived at nine LAL components. Although both Taylor (2013) and Kremmel and Harding (2020) have test developers among the main stakeholder groups, those cannot be equated to item writers. This is because test developers' responsibilities are in designing tests, including the test blueprint and item specifications, while item writers, according to the definition provided earlier in this section, are tasked with producing test items against existing specifications. This difference in roles leads to some differences in LAL needs, which was taken into consideration when designing the item-writing training course researched in the present study.

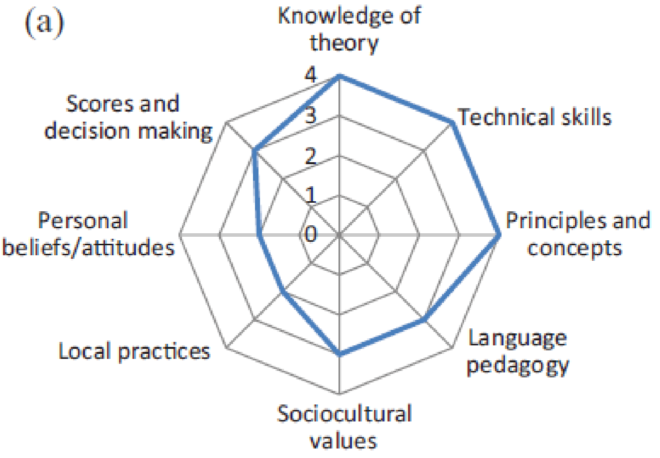


Figure 1-1. LAL profile for test writers (Taylor, 2013, p.410)

Cohort 1 of the training course was used to gain a better understanding of trainee item-writers' LAL needs. To this end, a questionnaire was administered to 18 trainees upon completion of the training (see Rossi, 2017), whose qualitative responses then served as a basis for developing a preliminary item writer LAL profile (*Figure 1-2*). The insights were used to inform course modifications for subsequent cohorts, with the course evolving with each run. A detailed overview of the training programme is provided in Section 3.2; the course syllabus can be found in *Appendix 1*.

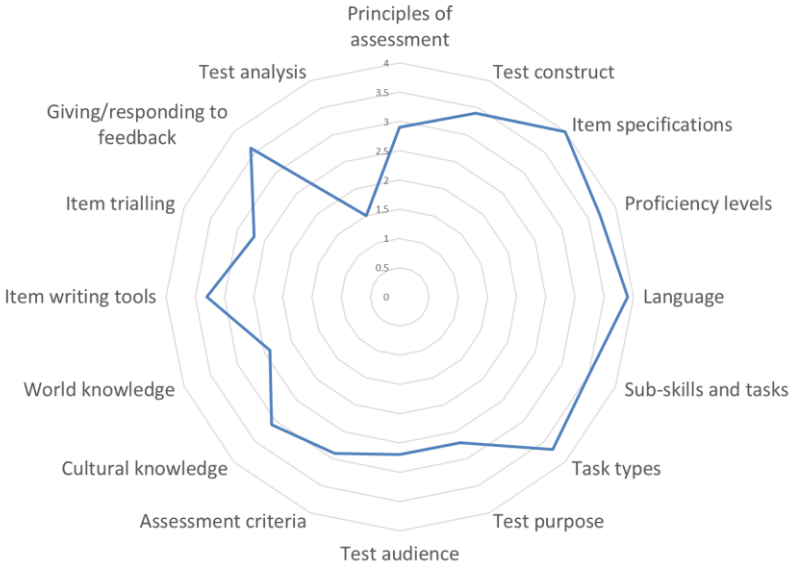


Figure 1-2. Item writer LAL profile (Rossi, 2017)

My assumption when starting this research project was that the item-writing training will result in a (more) skilled item-writing performance from the participants. Characteristics of skilled performance have been described in the literature. According to Welford (1968), it is "rapid and accurate" (p.12), while Proctor and Dutta (1995), summarising previous research into skill development, wrote that skilled behaviour is “goal-oriented and well-organised” (p.18) and is characterised by “strategies that enable efficient coordination of the various components of task performance” (p. 262), as well as with “a greater tendency to monitor one’s progress towards the goal” (p.243). Moreover, according to Proctor and Dutta (1995), experts are better able to evaluate their own performance. Based on these characteristics, my initial assumptions for the item-writing skill development resulting from the training course researched in this study were as follows: following the training, it would take participants less time to produce test items compared to before the training; the items written by a participant after the training would be of higher quality (i.e. would require little or no revision in order to be accepted for live testing) compared to the items written by the same participant before the training. After the training, participants would be better able to organise and monitor their own item-writing process; in particular, they would make use of some item-writing strategies, while before the training there would have been no evidence for strategy use; participants would also demonstrate better awareness of their own item-writing approach, compared to before the training. It was also expected that, following the training, participants would be better able to evaluate their own performance, for example they would be aware of any remaining deficiencies in their item-writing ability.

1.2.2 The theoretical framework for developing item specifications used for the training course

The item-writing course researched in this study trained participants in producing items against existing specifications. Sets of specifications were created for all practical item-writing activities carried out during the training, as well as for the two item-writing assignments: the pre-training assignment was used as a screening tool for course enrolment, while evaluations of the items produced for the post-training assignment were included in course completion certificates.

Although the training was provided by the British Council to its employees, it did not specifically train course participants in producing items for the Aptis test owned by the British Council (O’Sullivan et al., 2020), as the training was directed at enabling employees to join a variety of testing projects in the East Asia region. For this reason, item specifications used for the training were not those of the Aptis test. However, just like Aptis specifications, the specifications used for the training were informed by the principles formulated in the socio-cognitive framework for test development (O'Sullivan & Weir, 2011; Weir, 2005), as this framework primarily guides testing work conducted by the British Council. It was envisaged that those same principles would inform any test projects that the trainees might ultimately write items for.

In particular, in the training course, the models of reading (Khalifa & Weir, 2009) and of listening comprehension (Field, 2019), grounded in the cognitive processing approach, were used to define reading and listening sub-skills appropriate to target at different proficiency levels. For all item types, linguistic characteristics of items were defined in the specifications to make the items appropriate for the target test population stipulated in the training specs. Following Weir (2005), it was assumed that "[t]exts with more high-frequency vocabulary tend to be easier than texts with more low-frequency vocabulary" (p.77), therefore the lexical complexity of input and response for each item (see *Appendix 4*) were controlled by specifying their vocabulary frequency based on the British National Corpus-derived frequency lists. *Lextutor*[1] was used to generate item vocabulary profiles. It was also assumed that "[t]exts with less complex grammar tend to be easier than texts with more complex grammar" (Weir, 2005, p.78), therefore grammatical level of items was also controlled, with reference to the *Core Inventory for General English*[2]. Topics and communicative functions of item input and response were also controlled using the *Core Inventory*.

Following the socio-cognitive approach to test development, both situational and textual authenticity of items was afforded much attention in the specs (O'Sullivan, 2004; Wier, 2005). For example, writing was viewed "as a social act taking place in a specifiable context" (Weir, 2005, p.110); therefore, for writing prompts, the purpose of writing, the reader, and the response genre were to be specified by the item writer. The context of writing was to be established through an input message that served as the starting point for the expected response (see *Appendix 4*). Similarly, the listening task specifications required that input text characteristics (the genre, the speaker, the situation of speaking) were defined in the task instructions. With regard to textual authenticity, the listening task specifications stated that

[1] An online vocabulary profiler that classifies words of a text according to a vocabulary frequency list. The programme enables several corpus-based frequency lists, with BNC 1-20K (20 thousand most frequently used words in the British National Corpus) used for the item-writing purposes in this study https://www.lextutor.ca/vp/

[2] This document comprises "a core curriculum inventory for the English language based around key language points for each level, including grammar, vocabulary, discourse markers and functions" and is the product of a joint British Council–EAQUALS project to develop an English language teaching curriculum based on the CEFR https://www.teachingenglish.org.uk/article/british-council-eaquals-core-inventory-general-english

listening input texts were to “sound like authentic spoken English and not a written script read out” (*Appendix 4*), with spoken English characteristics defined after Carter and McCarthy (1997). The approach to listening task production, similar to that for the Aptis test (O’Sullivan et al., 2020), assumed that item writers produce both the text and a set of items to go with it.

Finally, test fairness concerns were also reflected in the specifications which stated that no specific background knowledge should be required from test-takers to produce a response to the writing prompt or to understand the listening input text. Moreover, item writers were to avoid topics that deal with religion, violence, abuse, controversial political issues or that might provoke negative emotions in test-takers. All item content had to be culturally unbiased and suitable for a general-purpose test. Grammar MC, writing prompt, and listening task specifications used for the pre/post item-writing assignment can be found in *Appendix 4.*

1.2.3 Principles of effective training that informed the item-writing course design

A constructivist approach to education (Steffe & Gale, 1995) informed the design of the item-writing training course researched in this study. In constructivism, learning is viewed not as mechanical transmission of general truths from teachers to passive learners but as a process that presupposes active learner involvement in practical activities. To this end, a large part of the item-writing training was dedicated to item-writing practice. Constructivism advocates learner co-operation whereby peers help each other in constructing their own knowledge (see, e.g., Zone of Proximal Development by Vygotsky, 1979). The item-writing course researched in this study was conducted online (see more on this in Section 3.2), with online discussions of language testing concepts, group analyses of test items, and collaborative item peer-feedback used as the ways to operationalise the constructivist idea of learner co-operation in an online environment (Mason, 2001).

Mayes (2001) offered a three-stage constructivist framework for online course design: (1) conceptualisation, when learners come to an initial understanding of a concept under review; (2) construction - “an activity in which the new understanding is brought to bear on a problem” (p.19); and (3) consolidation, which leads to full integration of the new understanding with the learners’ general framework of knowledge. The item-writing course researched in this study consisted of six two-week modules. Following the framework proposed by Mayes (2001), each module was structured in a similar way and consisted of (1) input on theoretical language testing principles and concepts relevant to the topic of the module (e.g. the construct and principles of assessing speaking); (2) a collaborative activity aimed at applying the said principles to the realities of item writing (e.g. analysing speaking prompts to identify whether they follow the language assessment principles introduced in the module); (3) an item-writing activity (e.g. producing speaking prompts against a set of specifications) followed with peer-feedback in small groups. Following the constructivist approach (Steffe & Gale, 1995), course tutors' role was seen as that of facilitators who introduced trainees to activities, guided the trainees through the item-writing process, clarified uncertainties, and provided feedback on trainees' items.

1.2.4 Item-writing trainees

Participants of the item-writing training course researched in this study were recruited among British Council China employees working primarily as language examiners. All of them held a university degree and a minimum of two years' experience teaching English as the second/foreign language. The course was designed to take this background into account. For example, because it was known that participants had experience teaching English to speakers of other languages, they were encouraged to capitalize on this experience when producing test items by, for example, deriving grammar MC item distractors from typical mistakes their students made. Because all participants also had experience living and working in at least one foreign country, the issues of cultural bias were discussed based on participants' own experience living in an unfamiliar culture.

The participants' primary motivation for taking the course was in diversifying their role within the organization by becoming involved in British Council assessment-related projects across the East Asia region. This role diversification was seen as desirable because the work of a full-time examiner was perceived as repetitive by some, while becoming involved in diverse item-writing projects promised some variety. Moreover, taking the item-writing course was seen as the first step towards further developmental and career promotion opportunities. Finally, some employees enrolled on the course with the view of increasing their employability outside the organisation. At the end of the training, those who fulfilled all course requirements received a British Council item-writing course completion certificate that specified the course syllabus, hours of instruction, and grades for the final item-writing assignment (see Section 3.2). At least some participants were hoping that having the certificate would improve their chances of becoming freelancers working as item writers for various language assessment bodies.

The above reasons might explain why, although course enrolment was completely voluntary and the time spent on the course was not remunerated, enough employees signed up to allow the run of three cohorts of the course, with 53 trainees completing the course, in total. Those trainees who demonstrated sound understanding of language assessment principles and good item writing ability upon completion of the training received promotion opportunities within and outside the organization; for example, some of them joined the British Council Assessment Solutions Team that "provides language assessment solutions for partners throughout East Asia including needs analysis, language assessment literacy training, test development, post-test services and teacher support" (https://www.britishcouncil.cn/en/exams/eaast-people).

1.3 Aims of the study

The goal of this study is to narrow the research gap identified in Section 1.1 by gaining empirical insights into item-writing skills and their development as a result of training. The study was conducted in the context of an existing induction item-writing training course, with the theoretical principles underlying the course outlined in the previous section. The research study based on the course adopted a mixed-method approach drawing on three sets of data: (1) expert judgements on the quality of items participants produced for the pre-training and post-training item-writing assignments; (2) interviews conducted with participants upon completion of their pre- and post-training assignments; and (3) feedback questionnaires administered to participants throughout the course. The study aims to achieve a better understanding of item-writing skills and their development as it happens during induction item-writing training. To this end, the study draws on two learning theories - Adaptive Control of Thought (ACT) and Communities of Practice (CoP)- to help interpret the study's findings. The theories were drawn on post-hoc: neither the study itself not the item-writing training course researched in the study were designed on the basis of these theories (see Section 2.3 for an outline of the theories' main precepts). The understanding gained in this study may result in practical suggestions for improving item-writing training effectiveness, which will increase item quality which, in turn, will help enhance the test validity argument.

1.4 Organisation of the thesis

This thesis consists of six chapters. Following this introduction, Chapter 2 reviews the literature on item writing and on skill development. The literature review covers both theory and empirical research into item writing, to clarify what has been explored so far and what research gaps still exist. A cognitive learning theory and a social learning theory are then presented, which are intended to help interpret the insights into item-writing skill development gained in this research project. At the end of this chapter, three research questions are formulated. Chapter 3 is devoted to the study's methodology. It provides information about the item-writing training course the study is based on, describes the overall research design, and details the two studies that form this research project: a Pretest-Posttest study and a Course Feedback study, both involving quantitative and qualitative data collection and analyses. Chapter 4 presents the findings associated with the three research questions. These include quantitative findings from expert evaluations of items produced by participants prior to and following the training, qualitative findings from interviews conducted with participants before and after the training, and quantitative/qualitative findings from four course evaluation questionnaires administered to participants throughout the training. Chapter 5 interprets and discusses these findings with reference to two learning theories – cognitive Adaptive Control of Thought - Rationale (ACT-R) theory (Anderson, 1993) and social Communities of Practice (CoP) theory (Lave & Wenger, 1991; Wenger, 1998). Finally, Chapter 6 includes a summary of the aims of the study and its key findings. It discusses theoretical, methodological, and practical implications of the study and indicates its limitations. The chapter concludes by considering the need for further research into item-writing skill development.

Chapter 2 Literature review

2.1 Introduction

This chapter is comprised of two parts reviewing literature from two different knowledge areas, both equally relevant to the current research project. Section 2.2 summarises literature on item writing and item-writing training, covering both theory and empirical research into item writing, with the aim to serve as a baseline for what has so far been done and what gaps still exist. Section 2.3 outlines two learning theories. While these are general theories that were not developed with item writing in mind, they are intended to help interpret the insights into item-writing skills and their development gained in this research project. Finally, Section 2.4 provides a brief summary of how the literature reviewed in this chapter has informed the current research project and proposes three research questions to be addressed by the study.

2.2 Item writing

In this section, the literature pertaining to item writing – primarily coming from language testing but also from the broader field of educational measurement - is discussed. Section 2.2.1 presents existing definitions of item writing, followed by a discussion of the role of item writing in test validation (2.2.2), item-writing documentation (2.2.3), item-writing procedures (2.3.4), item-writer characteristics and selection criteria (2.2.5), and item-writing training (2.2.6). Finally, previous empirical research into item writing and its relevance to the present research project is discussed (2.2.7).

2.2.1 Item writing: Definition

Davies et al. (1999) defined item writing as “the stage of test development in which test items are produced, according to a set of test specifications” (p.99), while Green (2014) called item writing “turning specifications into working assessments” (p.43) - the only definitions I was able to find in the language testing literature. Item writing has generally been regarded as ‘immature science’ in the literature (e.g., Cronbach, 1970; Nitko, 1984; Haladyna et al., 2002) and described by Haladyna et al. (2002) as “a loosely organized set of guidelines mainly transmitted via textbooks” (p.309). Existing item-writing guidelines are normally based on personal experience of expert item writers rather than on solid theoretical foundations: “Elder item-writers pass down to novices lists of rules and suggestions which they and their item-writing forefathers have learnt through the processes of applied art, empirical study, and practical experience” (Nitko, 1984, p.201). Regretfully, to date, insights into a theory underlying item writing, as well as empirical research into the practicalities of writing test items have not been published as extensively as for other areas of assessment. Shin (2012) contrasted the attention given in specialist literature to item writing and writer training with that given to rating and rater training, and concluded that the former “have not been properly introduced to testing communities, while issues related to rating and rater training have often appeared in language testing literature” (p.237).

2.2.2 Item writing and test validation

The scarcity of attention to item writing in the language testing literature is somewhat surprising, given that item writing directly impacts on test validity. Messick (1996) described six aspects of *construct validity* among which the content aspect deals with content relevance, representativeness, and what Messick (1996) called “technical quality” such as “appropriate reading level, unambiguous phrasing and correct keying” (p. 248). Weir (2005) referred to the content aspect of construct validity as *context validity* which includes such item characteristics as the clarity of rubrics and topic appropriateness. Moreover, Messick (1989) identified several threats to test validity that bear direct relevance to item writing: one is construct representation whereby the construct might be mis-, under- or over-represented in an item; the other threat is construct-irrelevant variance. The test validity aspects referred to by Messick (1989, 1996) and Weir (2005) directly pertain to item writing as they involve item quality, even though Messick and Weir themselves did not explicitly refer to item writing in their publications.

The educational measurement literature is slowly starting to recognise item writing as critical for valid assessment (see e.g., Haladyna & Rodriguez, 2013; Lane et al., 2015; Welch, 2006). For example, Haladyna and Rodriguez (2013) wrote that “the development of test items is an integral part of an argument for item validity” (p.7) and argued that test items should be a subject to validation as much as test scores are. Following Kane’s (2006) interpretative argument approach, Haladyna and Rodriguez (2013) proposed an argument-based approach to item validation with 16 questions to be answered while gathering validity evidence, including 'How are items developed?', 'Were items edited?', and 'Were items reviewed for fairness?' (p.12). This approach presupposes that efforts expanded during the item-writing stage of test development play an important part in building the overall test validity argument.

2.2.3 Item-writing documentation

The definitions of item writing provided by Davies et al. (1999) and Green (2014) suggest that any item writing should happen based on existing documentation. The documentation normally includes *models*, *test frameworks,* and *specifications,* which Fulcher and Davidson (2009) defined as follows: *models* are the most general documents "providing a theoretical overview of what we understand by what it means to know and use a language" (p.126), the Common European Framework of Reference (CEFR) being one of the most well-known of these frameworks in the European context. A *test framework* document states the purpose of a particular test and the test construct (normally selected from models). Fulcher and Davidson (2009) specified that a *Conceptual Assessment Framework* should consist of three parts: the test construct, test validity evidence, and the description of test tasks and items (pp.127-128). *Test specifications* are at the next layer of test documentation and outline details of each type of item and task earmarked for use in a test. In many cases, they are also the documents that item writers refer to whilst creating items and tasks.

Although live test specifications are normally confidential and difficult to obtain, from the samples published in the language testing literature (e.g. Davidson & Lynch, 2002; Luoma, 2004) one can conclude that specifications differ greatly in their content, presentation, and in the level of detail which test designers provide for item writers to work with. Davidson and Lynch (2002) argued that specifications should consist of: (1) a general description of what is to be tested; (2) the prompt attributes section that contains a detailed description of an item and task format and of what test takers will be asked to do; (3) the response attributes section that describes what test taker responses should contain; (4) the sample item(s). Alderson et al. (1995) believed that *specifications for test writers* should provide information about, among others, the test purpose, test taker population, target language situation, targeted language skills, intended tasks and items (pp. 11-14).

In language testing, some testing bodies create *item writer guidelines* based on their test specifications, which then become the primary documents item writers work with (see e.g., the *Standard Procedures for the Production of Listening Test Materials* by Cambridge Assessment in Geranpayeh & Taylor, 2013). Item writer guidelines normally include the level of practical detail unnecessary for test specifications for other stakeholders but essential for writing particular items, as exemplified in Elliott and Wilson (2013): "Item writer guidelines instruct writers that keys for constructed response tasks should not pose significant spelling problems" (p.180). Alderson (2010), while surveying Aviation English tests, concluded that both test specifications and item writer guidelines are useful in test production and called for "systematic procedures for review and editing of items and tasks to ensure that they match the test specifications and comply with item writer guidelines" (p.71). The usefulness of having both documents available to item writers is supported by Al-Lawati's study (2014) which found that item writers considered specifications and item writer guidelines two different documents: "the specs should cover the what, and the guidelines should cover the how" (p.145).

Even though there is no general agreement about the exact type and format of the documents item writers should be using, most assessment specialists agree that it is important to provide item writers with some documentation to refer to while writing items and tasks (Heaton, 1990; Hughes, 2003; Weir, 1995). This is especially true for large-scale standardized testing where there is a need "to control test tasks, so that new and equivalent versions of tests could be developed, trialed, and normed" (Davidson, 2012, p.198). However, how detailed specifications must be is an area of much debate. The main advantage of very detailed specifications is that they enable the production of highly comparable parallel test versions, thus increasing test validity (Haladyna, 2006). On the other hand, it is perceived by some that highly detailed specifications suppress item writer creativity and result in multiple clones of the sample item (Popham, 1994). Popham's solution is to provide item writers with "a set of varied, but not exhaustive, illustrative items" (pp.17-18). Davidson (2012) posed this debate as an issue pertaining test validity: "The finer grained the test specs are, the greater the control. The less grainy the specs, the greater freedom that the item writer has. What is the effect of this phenomenon on test validity?" (p.204). He concluded that more research is needed before testing science can offer some useful advice to item-writing practitioners on the issue.

2.2.4 Item-writing procedures

Downing and Haladyna (1997) argued for strict adherence to item-writing principles which should be adopted during the item-writing process. In their claim that item development is at the core of test validity, they believed that the item-writing process must be thoroughly documented, and evidence of compliance must be provided as part of the test validity argument. However, in reality the item-writing process is often afforded less attention compared to other practical aspects of operational testing, for example statistical analysis of item responses. This situation was criticised by Wesman (1971), who argued that statistical analysis can only help in identifying bad items while it cannot help in creating good ones.

It stands to reason that large testing organisations such as ETS or Cambridge Assessment would have well-organised procedures for item development. Pierce (1992) detailed test development procedures used at ETS during the then TOEFL Reading test production. At the time of publication, it included, first, commissioning freelance item writers to find suitable texts and write draft items based on detailed specifications. After initial items had been submitted, they were scrutinised by a member of a test development team and went through several cycles of revision that involved, besides six members of the test development team, a test specialist reviewer, the TOEFL coordinator, two editors and a sensitivity reviewer. Items that passed all reviews - while being heavily edited in the process - went for pre-testing with live TOEFL candidates, followed by item analysis to determine their item difficulty and discriminating ability. Finally, satisfactory items were included in one of the consequent TOEFL test forms.

IELTS item development procedures at Cambridge Assessment are available as *The IELTS Question Paper Production Process* brochure on takeielts.britishcouncil.org and are also mentioned in several IELTS research papers (e.g., Green & Hawkey, 2011). During the first commissioning stage, groups of trained item writers work from test specifications to produce test items. At the pre-editing stage, a meeting is held during which the materials are checked for the following characteristics: topic, topicality, level of language, suitability for the task, length, focus of text, style of writing, focus of task, level of task. Item writers then receive guidance on how to revise their items for resubmission. Revised items are submitted for editing; at editing meetings, texts and selected items are approved for pre-testing or are sent back to a writer for further revision, revised materials being re-edited at a subsequent meeting. Materials that pass the editing stage are sent for pre-testing with representative groups of language learners; pre-test responses are analysed using classical item statistics to evaluate the items' effectiveness. Finally, at a post-test review meeting, decisions are made on whether texts and items can be accepted for inclusion into potential live versions. Accepted items and tasks are then stored in an item bank to await test compilation.

From the two descriptions it becomes clear that item development is a complex and incremental process that occurs in a series of steps. Those steps often have to be reiterated, for example in multiple item revision sessions following feedback from item reviewers. It also seems clear that item writers are at the core of the whole process by choosing or creating suitable texts, writing and reviewing items, as well as sometimes doing item editing and reviewing work. Item writers might also combine several roles, acting both as writers of their own items and reviewers of their colleagues' items. Given the importance of item writers in operational language testing, it seems important to understand what the role and profile of item writers is, how item writers are selected and trained for the job, and how they carry out their work.

2.2.5 Item writers

Item writers' contribution to ensuring test validity is only starting to be acknowledged in the literature (Green & Hawkey, 2011; Rossi & Brunfaut, 2019). Green and Hawkey (2011) described the role of item writers within a large testing body as 'intermediate' because they have privileged access to the test provider while remaining external to it, and believed that this intermediate position "reflects the scope of the role that item writers normally play in the test production process"(p.112). The role is not only intermediary but also uncertain, with Shin (2012) lamenting that item writers are often viewed as quasi-professionals as compared to, for instance, test designers, raters, rater trainers, or data analysts. Shin (2012) asked questions that remain largely unanswered within the scholarly literature: "Who are the individuals who develop test items? And who trains them?" (p.242).

Ebel (1963) listed five requirements for a good item writer: thorough mastery of the subject matter, well-developed educational values, psychological and educational understanding of the test-taking population, mastery of verbal communication, and knowledge of item-writing techniques. Wesman (1971) reiterated the five requirements while adding a sixth one – specialization. He believed that item writing is not a unitary skill; an item writer may be proficient at writing vocabulary items but unable to construct good reading comprehension tasks.

In the context of language testing, Alderson et al. (1995) believed that item writers should be "experienced teachers of similar students or relevant subject areas" (p.40) but stressed that this alone is not enough as a good item writer will possess creativity, sensitivity, insight and imagination - the four elusive qualities that "are difficult to define and difficult to identify in prospective item writers, but very obviously missing in poor item writers" (p.41). The ILTA Guidelines for Practice (2007) emphasize the necessity for item writers to be "well versed in current language testing theory and practice" (p.3), while the EALTA Guidelines (2006) include considerations of relevant teaching experience and training (p.3).

Large testing bodies have their own selection criteria for hiring item writers. Ingham (2008) provided some insight into Minimum Professional Requirements (MPRs) for item writers at Cambridge ESOL. They include a degree, an ESOL qualification, and five years' teaching experience. "Some familiarity with materials production is also required, as is some involvement in preparing students for Cambridge ESOL examinations; writing and publishing experience is also desirable" (Ingham, 2008, p.6). Alderson et al. (1995) conducted a survey of EFL examination boards and found that the boards varied in their requirements to appointing item writers. Some put emphasis on appropriate qualifications, such as a university degree or EFL/ESL qualifications. Others asked for teaching / examining experience, or experience in the relevant subject area (Alderson et al., 1995, p.65).

Good item writers are in short supply (Buck, 2009) and, to complement a set of characteristics that cannot always be found in one person, teams of item writers are sometimes employed, especially in Language for Specific Purposes (LSP) testing where language testers collaborate with subject specialists, or when tests of less-commonly taught languages are written by assessment experts together with so-called "language informants" (Ryan & Brunfaut, 2016).

Two recent empirical studies yielded some insight into item writer characteristics. Salisbury (2005) used verbal protocol methodology and a theoretical framework for the study of expertise in her PhD research to explore how item writers produce listening comprehension tasks. She found that two characteristics are vital for item-writing expertise: previous experience in test production and "contact with the target domain, in the form of preparing students for such tests" (Salisbury, 2005, p.286). Her empirical findings also corroborate Wesman's (1971) idea that item writing is not a unitary skill. In Salisbury's study, several participants "exhibited a 'jagged profile' of skills, and even the 'top scoring' experts demonstrated relative weaknesses in some areas" (p.287). Contrary to Ebel (1963) who claimed that item writers are not born but made through hard work and experience, Salisbury (2005) claimed that several pre-existing characteristics contributed to the listening item-writing quality in individual writers, among them "an ear for 'speakerly text' and the ability to create it from a base [sic] of a written text" and "particularly efficient aural memory - both working and long-term" (p.293).

Kim et al. (2010) reported on a case study of item-writing processes when producing grammar and reading items for a language test. They found that the native/non-native distinction makes a difference, as non-native English speakers in the study felt under pressure in terms of the linguistic accuracy of the items they produced. At the same time, being non-native speakers of English offered those participants some advantage as they had previous experience taking ESL/EFL proficiency tests, which enabled them "to better perceive how test takers would respond to the items because they can also look at test items from the test takers' point of view" (Kim et al., 2010, p.171). The study participants also mentioned that being focussed and striving for perfection in their work were two personal characteristics that helped them during item writing.

2.2.6 Item-writing training

Shin's (2012) questions "…who are the individuals who develop test items? And who trains them?" (p.242) sound largely rhetorical. A general impression from reading the literature is that the undefined 'experienced item writers' are expected to serve as mentors to novices, and the training largely to happen in employment. As Ebel (1963) put it, item writers "must usually learn to write by writing" (p.188), and the situation fully applies to language testing. In fact, Hughes (2003) described stages of language test development, emphasizing that "all staff who will be involved in the test process should be trained" (p.66), but not including item writers among those to be trained, having omitted them from the list of interviewers, raters, scorers, computer operators and invigilators. In a survey of Aviation English tests, Alderson (2010) found that half of the surveyed testing organisations did not provide any training for their item writers. One explanation might be that item writer training is seen as expensive, therefore testing organisations are unwilling to invest large sums in it hoping that, if they hire professionals with suitable qualifications and experience, those item-writing skills will develop in due course.

At the same time, it has been argued in the educational measurement literature that conducting item-writing training "constitutes evidence for item validation" (Haladyna & Rodriguez, 2013, p.22) because untrained novice item writers tend to produce poor-quality, flawed, idiosyncratic items. Downing (2006) urged that all those who have responsibility for writing any kinds of test item should go through formal training. Vaughn (1963) stated that item-writing training is not only desirable but essential, and enumerated what the training organisations must consider when preparing such training, including the scope of training, amount of time devoted to each item-writing technique, mode of training, and training approaches.

The necessity of item-writing training has been confirmed through empirical studies in the field of educational measurement for medical science. For example, Jozefowicz et al. (2002) analysed 555 in-house medical school examination MCQs on a five-point quality scale and found that MCQs written by faculty trained in item writing had a mean score of 4.24, while questions written by faculty without formal item-writing training had a mean score of 2.03. They concluded that the in-house medical school examination materials were of relatively low quality and suggested that the quality of examination questions could be significantly improved by formally training question writers.

However, educational measurement specialists rarely go further than simply acknowledging the importance of item-writing training. Rare exceptions are publications by Welch (2006) and Haladyna and Rodriguez (2013) who provided practical recommendations on how to organise such training. Welch (2006) distinguished between three major training modes: online, mail-out and face-to-face. The first training mode has the advantages of any online training: it can involve larger numbers of trainees, does not incur high travel and accommodation costs, and participants can work according to their own schedule (p.308). A simplified version of online training is the mail-out approach whereby participants work through written training materials sent to them via email. Face-to-face workshops, on the other hand, offer participants a valuable opportunity to come together as a group, and discuss and try out their ideas together. Welch also argued that "workshops have the advantage of immediate feedback on the quality of produced items" (p.308). As for the training length, Haladyna and Rodriguez (2013) believed that item-writing training can last from several hours to several days, depending on training needs.

In terms of training approaches, Downing (2006) suggested that a hands-on training workshop should be structured as an "instruction - practice - feedback - reinforcement loop" (p.11).

Welch (2006) proposed the following agenda for a face-to-face item-writing training workshop on writing prompts for performance assessment: (1) discussing the purpose and audience of the assessment; (2) presentation of test specifications and test development process; (3) general guidelines for prompt writing, such as sources, copyright issues; (4) presentation of the prompt templates or 'item shells'; (5) presentation and discussion of successful and unsuccessful prompts; (6) trainee prompt writers generate topics for consideration followed by the topic discussion and approval; (7) trainee prompt writers create prompts from the approved topics (p.309). The outline for an item-writing training session proposed by Haladyna and Rodriguez (2013, p.22) largely reiterates Welch's (2006) suggestions. Importantly, both schedules include item-writing practice and group discussions of items. Haladyna and Rodriguez (2013) emphasized the latter because "'[t]o hear colleagues discuss your item and offer constructive advice is valuable both for improving the item and for learning how to write better items" (p.23).

Al-Lawati's (2014) unpublished doctoral dissertation provided some rare valuable insights into item-writers' own perceptions of their training needs. Item writers who took part in a focus group were asked whether training events they attended were beneficial and whether they had any suggestions for future training events. All focus group participants felt that formal training in item writing was useful and necessary, and expressed the wish to attend more training events. They also suggested topics for future training, including 'feedback on their items', 'sample items', 'sources of good texts', 'interpretation of topics', 'CEFR levels and scales', and 'collaboration' (Al-Lawati, 2014, pp.155-157).

In the field of language testing, item-writing training is typically mentioned in association with prominent testing bodies who have their own training approaches and procedures. For example, Ingham (2008) and de Jong (2008) offered some insight into how item-writing training was conducted (and might still be) at Cambridge Assessment and Pearson Education, respectively. A generic training weekend that serves as an induction training for item writers at Cambridge Assessment would normally involve: (1) an overview of Cambridge Assessment examinations and an introduction to the principles of test design and production as well as the basic terminology used to describe test questions; (2) two-hour sessions on the techniques of writing particular item types, including input from the trainer and group activities drawing on the ideas and experience of the participants; (3) an overview of writing for particular skills papers where participants were introduced to how writing for each of these skills had an impact on the item type and any implications for the item writer; and (4) text selection and adaptation (Ingham, 2008, pp.6-7). Item-writing training at Pearson Education, as explained by de Jong at the 2008 EALTA conference, is/was organised as a one-day face-to-face workshop that covers the following agenda: introduction to the CEFR and practice with scale descriptors, selecting texts, technical item-writing principles, sensitivity issues, working with item templates, item reviewing, feedback on acceptance rate and reasons for rejection (de Jong, 2008).

2.2.7 Empirical research into item writing

This section draws on recent empirical research related to item writing and conducted in the fields of language testing and educational measurement. One group of studies investigated the effectiveness of item-writing training (2.2.7.1), while the other group looked into the processes experienced and/or inexperienced item writers engage in while producing test items (2.2.7.2).

2.2.7.1 Item-writing training effectiveness

Although the training practices described in Section 2.2.6 provide valuable insights into item-writing training approaches of large language testing bodies, no empirical proof of the training effectiveness, to the best of my knowledge, has been presented in the language testing literature. A wider literature search identified nine recent (quasi-)experimental studies evaluating the effectiveness of item-writing training in the fields of medical education, dentistry, and school maths education – see *Appendix 15*.

In all nine studies, the training was conducted for teaching staff with item-writing responsibilities, rather than for professional item writers. The training mainly focussed on writing MCQs; additionally, the training for medical school faculty reported in Naeem et al. (2012) included short-answer questions and objective structured clinical examination checklists, while the training for school maths teachers described in Yurdakul et al. (2020) included open-ended and true/false questions. The training approaches varied considerably, for example a 30-min online presentation followed by a 10-min Q&A session (Scott et al., 2019), three 3-hr workshops conducted over the period of several months (Iramaneerat, 2012), or a one-week full-time training event (Naeem et al., 2012). In Hamamoto Filho & Bicudo (2020), the intervention took the form of feedback on items produced in the previous year.

Irrespective of the intervention type, all studies reported a significant improvement in the quality of items produced after the training. One of three methods was used to evaluate the training effectiveness: (1) item evaluations by human judges against a rating scale; (2) item functioning in a live test, such as item difficulty, discrimination, number of functioning distractors, and student performance; (3) participants' feedback on the training. Three studies (Iramaneerat, 2012; Tricio et al., 2018; Yurdakul et al., 2020) used a combination of two methods.

The 'item evaluation' method (1) was used in seven of the nine studies. For such studies to obtain meaningful results, the judges must be suitable for their role, the rating scale must allow for valid evaluations, and the rating process must strive to eliminate potential judges' subjectivity. Unfortunately, each study's methodology had at least one problem:

- no information was provided about the number of judges in the study or the way the judgements were made (Gupta et al., 2020; Hamamoto Filho & Bicudo, 2020; Tricio et al., 2018);
- the judges' credentials were not established (Dellinges & Curtis, 2017; Scott et al., 2019);
- the evaluation was not blinded; for example, in Naheem et al. (2012) the only judge, who was also the trainer and the researcher, knew whether the items came from before or after the training;
- two judges evaluated the quality of post-test items only, but it is unclear how discrepancies in the judgements were resolved (Yurdakul et al., 2020);
- the rating scale used to evaluate the quality of MCQs did not include important aspects of MCQ quality (Dellinges & Curtis, 2017; Scott et al., 2019);
- two or more different aspects of item quality were conflated into a single criterion (Dellinges & Curtis, 2017; Scott et al., 2019).

To illustrate the last point, one criterion in Dellinges and Curtis (2017) read: "A single clearly formulated problem in simple language is presented in the stem of the item. As much of the wording as possible is in the stem" (p.950). This criterion conflates five (!) different requirements: there is only one problem per item; the problem is clearly formulated; the language is simple; the problem is presented in the stem; as much of the wording as possible is in the stem. Because the items were judged on a two-level 'yes/no' scale, a 'no' for only one or for all aspects of the criterion would result in the same '0' score, thus making the item evaluations highly imprecise.

When the 'item functioning analysis' method (2) was used, the way it was applied could have potentially invalidated the study results. In two studies (Abdulghani et al., 2015; Tricio et al., 2018), tests as a whole were analysed irrespective of who contributed items for the test - trained or untrained staff. In Abdulghani et al. (2015), all items produced in the post-training year went through quality review and editing, but it is unclear whether the same procedure was followed in the pre-training year. These two studies (Abdulghani et al., 2015; Tricio et al., 2018) also used students' test performance as a proof of training effectiveness; however, because the tests were administered to different student cohorts and one year apart, there might not be a correlation between test-takers' performance and item quality, especially as no methods of score equation (e.g., anchor items or Rasch analysis) were reported. Moreover, both an increase (Abdulghani et al., 2015) and a decrease (Tricio et al., 2018) in students' mean scores were interpreted as evidence of training effectiveness.

It also seems that, for some of these studies, the claims of training effectiveness may have been inflated. For instance, Yurdakul et al. (2020) wrote that "the participants made progress" (p.98) while their posttest-only study design with participants who had had previous item-writing experience does not allow for such claims to be made. Also, the way the data was collected and analysed in some studies might have helped to make the training effectiveness seem larger than it really was. For instance, in Dellinges and Curtis (2017), and Naeem et al. (2012) no new items were created after the training, but the participants were asked to make improvements to the pre-training items. Moreover, in Naeem et al. (2012) the items went through two rounds of focussed peer- and trainer-feedback. In several studies, pre- and posttest data on item quality or item functioning was compared for the item-writers' cohort as a whole and not for individual trainees (Hamamoto Filho & Bicudo, 2020; Iramaneerat, 2012; Naeem et al., 2012); only the sums of scores for each item were considered and not the scores on individual criteria (Dellinges & Curtis, 2017; Scott et al., 2019). This approach to data analysis also made the studies less informative because it was impossible to determine the effect of training on individual participants as well as on individual aspects of item quality, which renders the studies' results irrelevant for research into item-writing skill development.

Overall, after close consideration of the nine studies' methodology, their claims on the item-writing training's effectiveness do not seem fully convincing. Moreover, because the studies employed vastly different approaches to item-writing training, all of which, reportedly, resulted in improved item quality, it is impossible to determine which training approaches were more beneficial. This makes it difficult to draw implications for item-writing practice, apart from a very general observation that any training is better than none, which seems, in fact, self-evident. Finally, multiple methodological flaws in the data collection and analyses point to the fact that the methodology for research into item-writing training effectiveness is still in its infancy.

In fact, Gupta et al. (2020) took a more measured stance in interpreting their results, compared to the other eight papers discussed above. Gupta et al. (2020) admitted that, although the statistical analysis of item evaluations resulted in overall significant improvement, the improvement "was not sufficient to have an educational impact" (p.212) because it was limited to better homogeneity of options and did not manifest in other aspects of item quality; moreover, no significant improvement was observed in terms of item facility values, discrimination, or non-functioning distractors. The authors attributed the limited training impact to the way the training was organised (one 3-hr input session), advocating for a longitudinal training programme that would incorporate practical item-writing events, group discussions, and peer feedback.

2.2.7.2 Item-writing processes

It stands to reason that any training aimed at developing item-writing skills would want the trainees to follow those item-writing processes and adopt those item-writing strategies that result in high-quality items. Importantly, such processes and strategies should be known not only anecdotally through individual item-writer's experience but backed by empirical research. However, similar to the research into item-writing training effectiveness, very few studies have investigated the process of writing test items or what constitutes an expert item-writing performance. To the best of my knowledge, only five studies within the field of ESL/EFL have been conducted so far. In the following paragraphs, a brief summary of each study's findings is provided.

Kim et al. (2010) conducted a case study with four inexperienced item writers who worked as a group to develop a set of grammar items and of reading items. The data collection involved item writers keeping a reflective journal, interviews, surveys, and document analysis. The study was limited to three aspects of the item-writing process: item-writers' use of item specifications, effects of group dynamics, and effects of individual item-writer characteristics such as experience, language background, and personality. The study revealed that novice item writers were reluctant to use the specifications and preferred to rely on their previous classroom test development experience. The item writing itself was "both a personal and group process" (p.147) with the item-writers' personalities and the item writers being either native or non-native English speakers affecting the group dynamics and the nature of peer-feedback on the items.

Green and Hawkey (2011, 2012) conducted a case study that involved four experienced and three inexperienced item writers selecting and adapting reading texts for the IELTS Reading Paper. The data was collected through focus groups, individual interviews, and observation of an editing meeting. The study participants also produced "flow charts of their writing process" (Green & Hawkey, 2011, p.115). The study highlighted the complexity of the item-writing process and the necessity for "intensive individual and collective work" (Green & Hawkey, 2011, p. 126) in producing language test items. The authors concluded that item writing requires a high level of "expertise and application" (Green & Hawkey, 2011, p.127).

For her doctoral research project, Salisbury (2005) carried out two studies – one exploratory (with six experienced item writers, using interviews) and one quasi-experimental (with five experienced and five inexperienced item writers, using verbal protocols) - to investigate expert performance and to compare it with novices' performance in producing listening tests. Drawing from multiple theoretical frameworks on the nature of expertise, Salisbury (2005) concluded that experienced item writers have "a predictable core of domain knowledge - declarative, procedural and strategic" (p.295) but there is also "considerable individual variation in performance and acquisition processes" (p.295). The findings indicated that listening item-writing skills are specialist, manifold, but also distributed in that "individual item writers seldom exhibit expertise in all aspects of domain practice, and need to work as part of complex domain system [sic] in order to bring their task to completion" (p.295).

Ho (2019) conducted a study into "the development of language assessment literacy of pre-service ESL teachers through the processes of item writing" (p.1) for his Master's degree. Seven novice item writers produced prompts for an integrated-skills placement test, with the study's data including group discussions, individual interviews, and multiple item drafts. The findings were largely discussed from a genre theory perspective, but the author also highlighted the importance of collaboration in the item-writing process, claiming that participation in group discussions and acting on peer feedback resulted in learning about item-writing.

Additionally, there is a preliminary report on an ongoing study conducted by Ngo (2016) involving one listening item writer for a new CEFR-aligned high-stakes test in Vietnam, with data collected via narrative frames, verbal protocol, and reflective journals. Ngo (2016) used Cultural Historical Activity Theory (CHAT) as a framework for exploring the factors that mediate the item-writing activity and found that the activity was mediated by the item writer's educational background, previous working experience, item-writing training, and the practice gained during his item-writing work.

Because so few studies into the item-writing process have been conducted, I additionally consulted fields other than language testing. The search revealed two such research projects: one conducted in the USA by Dennis Fulkerson and colleagues in the field of school science education, and the other in the UK by Martin Johnson and colleagues, who researched item writing for GCSE tests in Biology, Geography, Mathematics, and Physics.

The first research project aimed to investigate item-writing expertise based on the Theory of Insight Problem Solving. Three consecutive studies were conducted, the first of which (Fulkerson et al., 2009) used verbal protocol analysis (VPA) to investigate cognitive processes of expert item writers while producing MCQs for a science test. It was found that the item-writing process in the study consisted of three phases: (1) *representation* when the item writers created a mental model of the item-writing task; (2) *exploration*, when the item writers looked for content to produce the item; (3) *solution* when the item writers completed the item "by finding a workable solution that satisfies the predetermined constraints" (Fulkerson & Nichols, 2010, p.3). The second study (Fulkerson et al., 2010) used VPA to compare the cognitive processes of two experienced vs. one inexperienced item writer. It was found that the inexperienced item writer spent longer defining the problem and demonstrated "frequent stalled or backward movement in the problem space" (p.15). The cognition of the experienced item writers, on the other hand, quickly moved forward through the problem space. In the last study (Fulkerson et al., 2011), the cognitive processes of four novice and five expert item writers were investigated using the same methodology, with the aim of determining the role of knowledge structures in the item-writing process. It was found that, for both novice and experienced item writers, assessment content knowledge and general item-writing knowledge were primary for the creation of quality items, while domain-specific content knowledge and pedagogical knowledge were secondary.

The second research project (Johnson et al., 2017) involved seven professional item writers who were video-observed while producing GCSE test items, with stimulated-recall interviews conducted after the item-writing session. It was found that the cognitive item-writing process consisted of three phases*:* (1) *thinking about writing,* (2) *writing and reflective thinking*, and (3) *reviewing*. Besides the cognitive perspective, the researchers also posited item writing as a social act situated within the professional testing community: item-writing resources served as important artefacts of the item-writing process, while the item writers were attempting to adopt test-takers' and item-reviewers' perspectives to conform to the expectations of the community. This social perspective was further developed in Constantinou et al. (2018) who analysed the data from Johnson et al. (2017) using Bakhtin's concept of 'multivoicedness' to identify the various voices involved in informing and shaping the item-writing process. The authors produced a macro model of test writing comprised of two overlapping voices – authoritative word (official discourse) and internally persuasive word (item writer's personal beliefs) - enclosed within the item writers' community of practice which is, in turn, enclosed within the society with its prevailing ideologies of the 'politics of knowledge' and 'fairness'.

Although all studies outlined above produced novel and interesting findings, they have a number of limitations. A major limitation is the small number of participants in each study, which does not allow for generalisation in the findings. Different research foci also prohibit the generalisation of findings from different studies. For example, within the field of language testing, Green and Hawkey (2011, 2012) looked into reading text selection and adaptation, Kim et al. (2010) investigated the use of specifications, group dynamics and individual item-writer characteristics, while Salisbury (2005) compared the item-writing processes of experts and novices. Authors' definition of an item-writing expert is also unclear. In most studies, expertise is related to the number of years in service, but the exact years differ from study to study, for instance anything over 1 year (Fulkerson et al., 2010) or between 3 and 27 years (Johnson et al., 2017). Green and Hawkey (2011) related expertise to the number of successful item-writing commissions, but the range is, again, rather wide – from seven to 25. Salisbury (2005) pointed out that it is not the number of years but the ability to produce high-quality items that should define expertise; however, she herself categorised her participants according to the number of years in service.

The studies by Green and Hawkey (2011, 2012) and Kim et al. (2010) are framed as experience-sharing and lack an underlying analytical framework to interpret the findings. Salisbury (2005) based her research on the notions of expertise existing in the literature; however, because she included many different frameworks in the discussion, she was only able to compare individual findings with one framework at a time, without offering a coherent picture. The research project by Fulkerson and colleagues (2009; 2010; 2011), on the contrary, is firmly grounded in theory: they adopted a cognitive perspective using the Theory of Insight Problem Solving as a framework for analysing the item-writing process. This enabled them to articulate the findings in a coherent manner but, as noted by Johnson et al. (2017), the framework also limited their ability to interpret the data in that they "treated question writing as an internal process governed by cognitive and psychological mechanisms" (p.704) and overlooked its social dimension. Consequently, Johnson et al. (2017) adopted a dual cognitive and social outlook (the latter was based on the Communities of Practice theory), while in a subsequent publication based on the same study Constantinou et al. (2018) used Bakhtin's concept of 'multivoicedness' as the foundation for their macro model of test writing. The model, however, is interesting only as a theoretical construct because the authors failed to draw implications for item-writing training.

The most significant reason why the above studies have rather limited relevance to the present research project, however, is in their static view on expertise. The studies investigated the item-writing processes of either inexperienced (Ho, 2019; Kim et al., 2010) or experienced (Johnson et al., 2017) writers, or compared the processes of experienced and inexperienced writers (Fulkerson et al., 2010; Green & Hawkey, 2012; Salisbury, 2005) at one point in time. To the best of my knowledge, no longitudinal research has so far investigated how item-writing approaches evolve with the development of expertise and as a result of training, something which is the primary concern of this research project.

2.3 Theories of learning: individual (cognitive) and social (situated) perspectives

The present study is concerned with training people to develop their item-writing skills, and the two theories of learning outlined in this section are intended to help interpret the insights into item-writing skill development gained in this research project. Human learning is a multidimensional activity. Many different types of learning theories have been proposed, each emphasizing a different aspect of learning. Behaviourist, cognitive, and constructivist learning theories, all originating from psychology, are concerned with individual learning and the mental processes behind it. They look at learning as an individual activity independent of social interaction. Lantolf and Thorne (2006) explained this phenomenon:

> Wilhelm Wundt, considered by many to be the founder of modern psychology... needed to formulate a stable object of study that was somehow also independent of

> people as social beings. His solution was to abandon the embedded, and necessarily unstable, qualities of human mental processes and assign these to fields such as anthropology... Once this bifurcation was institutionalised, it was taken for granted that psychology could then concentrate on what were assumed to be the stable and universal features of the mind (p.152)

Various social learning theories originating from sociology, pedagogy, anthropology, and human resource management have adopted a markedly different perspective on learning by situating it within human society – they claim that all learning is context-bound and can happen only through interaction with other people and material environments. There now exist many social learning theories - e.g. Activity theory, Socialization theory, Organizational theory - all of which emphasize the collaborative nature of learning and claim that learning potential increases in a community environment.

This study has drawn on two learning theories - Adaptive Control of Thought (ACT) and Communities of Practice (CoP) - to help interpret the insights into item-writing skills and their development as they happened in the item-writing training course. It is important to emphasize, however, that the theories were drawn on post-hoc: neither the item-writing training course, nor the study itself were designed on the basis of these theories. Instead, as elaborated in Section 1.2, the item-writing training was informed by the socio-cognitive framework for test development, the scope of the training was defined based on the current understanding of LAL expected of item writers, and the pedagogical principles were informed by the constructivist approach to online education. The study's design, including methods for data collection (detailed in Chapter 3), similarly were not informed by the two learning theories; the theories were drawn on after the data collection was finished, during the process of data analysis and interpretation. The decision to use the theories is explained in more detail in the paragraphs that follow.

After I collected and analysed the data, I discovered that the quantitative and qualitative study findings did not present a simple, linear picture of skill development, suggesting that item-writing skill development might be more complex than initially expected. Looking for possible explanations for the findings, and given that item-writing is a skill (Section 1.2.1), I turned to theories of skill acquisition. Among the skill acquisition theories I considered were Skill Theory by Fischer (1980), Instructional-Design Theory for skill development by Romiszowski (2009), and several theories of expertise (e.g. Chi, Glaser & Farr, 1988; Ericsson & Charness, 1997). Although these theories offered some useful insights and could be applicable to some of the study's findings, for example the notion of a skills cycle proposed by Romiszowski (2009) or the developmental range concept proposed by Fischer (1980), I felt that the theories did not provide the affordances to explain the study's findings in their entirety. At the same time, I believed that drawing on a different theory to explain each individual finding, as was done by Salisbury (2005) for her doctoral study, would be unsatisfactory in the case of the present study: I was hoping to arrive at a coherent and comprehensive model of item-writing skill development, while drawing upon many different theories would make the discussion fragmented.

During my explorations, I was particularly drawn to the Skill Acquisition Theory developed by DeKeyser (2007). As acknowledged by DeKeyser himself, Skill Acquisition Theory, which is used to explain the process of second language acquisition, originates from works by John Anderson, who developed the Adaptive Control of Thought (ACT) theory – a general theory of cognition that can be applied to the development of any cognitive skill. In the belief that it is meaningful to draw ideas from the original source, I turned to Anderson's work (1981, 1993, 1996, 2010) and felt that the ACT theory might provide a suitable framework for explaining the findings of this doctoral study. I am fully aware, though, that any theory is neither wrong nor right in itself - it can only provide affordances for interpreting research findings. In my opinion as the researcher, the ACT theory provided the best affordances for explaining this study's findings.

A preliminary discussion of this study's findings was presented at the 41st Language Testing Research Colloquium in Atlanta (Rossi & Brunfaut, 2019) and provoked a meaningful discussion. One comment that was made was that, although the item-writing skill development process as hypothesised in the presentation was very promising, it did not embrace the whole of the item-writing skill development because it overlooked its social dimension. Communities of Practice (Wenger, 1998) and Sociocultural Theory (Lantolf & Thorne, 2006) were suggested as potential theories to explain the social aspect of item-writing skill development. It is important to note that the commenters did not reject the idea of interpreting the findings from the individual cognitive perspective; rather, they suggested that the perspective should be complemented with a social view on skill development.

Subsequently, I researched a range of social learning theories, including Sociocultural Theory (Lantolf & Thorn, 2006), Cultural-Historical Activity Theory (Engestrom, 2014), Organizational Theory (Argyris & Schon, 1978), and Communities of Practice Theory (Wenger, 1998). Based on my insights into this study's data, it seemed to me that the CoP might offer the best affordances to explaining the social nature of item-writing skill development. I also suspected that combining the ACT theory which views skill development as an individual, cognitive process, with the CoP theory which views skill development as a social situated activity, might result in a comprehensive account of the item-writing skill development as it happened during the item-writing training course researched in this study. Drawing on these two theories to interpret this study's findings was a matter of my choice as the researcher, informed by my knowledge of the data. While it may not be the only possible choice, I was hoping that it would result in a useful, although by no means definitive, exploration of item-writing skill development.

Combining two theories of learning in one study is not unique. Although each individual learning theory offers a useful perspective, none of them can claim to explain learning in its entirety, something that has been recognised by the authors of the theories themselves. For example, Etienne Wenger, the author (together with Jean Lave) of the Communities of Practice theory, wrote in his book *Communities of practice: Learning, meaning and identity* (1998): "I am not claiming that a social perspective of the sort proposed here says everything there is to say about learning… Nor do I make any sweeping claim that the assumptions that underlie my approach are incompatible with those of other theories" (p.279). Indeed, it has long been felt that approaching human learning from only one perspective might prove insufficient. For example, in the field of human resource development (HRD) concerned with adult learning within organisations, learning is traditionally approached from two perspectives – as an individual and as a social process – both of which are seen as equally important (see, e.g., Kirwan, 2013). There have also been attempts to propose a holistic learning theory (Yang, 2004) that would combine the individual and social dimensions of learning because "most of the existing adult learning theories tend to narrowly define knowledge and learning and thus fail to offer adequate explanation for adult learning" (Yang, 2004, p.260).

Some of the studies into item writing reviewed in Section 2.2.7.2 of this chapter also drew on multiple theories to explain their findings. For example, Salisbury (2005) used three different models of skill acquisition to discuss the findings of her doctoral study into listening item writing: "the information processing model (IPM) of cognitive psychology, Dreyfus and Dreyfus's proceduralisation model, and Bereiter and Scardamalia's continuous process theory" (p.65). It should be noted, though, that all three models are concerned with the individual cognitive dimension of learning, therefore they can be seen as competing rather than complementing each other. In another example, Fulkerson et al. (2009, 2010) viewed item writing as an individual cognitive activity seen through the prism of the Theory of Insight Problem Solving. However, this approach was criticised by Johnson et al. (2017) as limiting and unidimensional. Johnson et al. (2017) advocated the necessity of a “broader approach to examining the process of question writing, one that views question writing not merely as a cognitive process but as a socio-cognitive phenomenon” (p.704). Consequently, in their study into professional practices of seven experienced item writers, Johnson et al. (2017) drew on two theories - the cognitive model of writing proposed by Flower and Hayes (1981) and the CoP theory proposed by Wenger (1998) - to discuss the findings. The present research study also aims to introduce a balanced view on item-writing skill development by adopting a dual - cognitive and social - perspective. It takes the Adaptive Control of Thought (ACT) theory as the framework for the cognitive side of item-writing skill development, while the Community of Practice (CoP) theory is used to account for its social nature.

The ACT theory of skill acquisition, developed by John Anderson (1993), is “the most widely used cognitive architecture in cognitive science” (Anderson, 2010, p.v) that has been advocated for a unified theory of cognition (see e.g., Newell, 2013). ACT is primarily a theory of human cognition of which the theory of skill acquisition is a part and is “one of the most influential theories of skill acquisition and, to date, the most comprehensive” (Speelman & Kirsner, 2005, p.40). Its claims are backed by a large body of empirical evidence gained through research into the acquisition of text editing (Anderson & Singley, 1993; Singley & Anderson, 1985), computer programming (Anderson, 1987; Anderson et al., 1993b), problem-solving (Anderson et al., 1993a) and other complex cognitive skills. ACT is based on the notion of two types of knowledge – declarative and procedural - and offers a comprehensive model of skill acquisition that accounts for the acquisition process from its initial stage to reaching an expert status. Moreover, ACT accounts for the role of training in skill acquisition. All the above features made the theory attractive for research in many areas of science and humanities. For example, it has been used in neuroscience to research brain activation in brain imaging (MRI) experiments (see e.g., Sohn et al., 2003; Qin et al., 2003). ACT has also been used to model human behaviour when performing complex tasks such as driving (Salvucci, 2006) and flying (Byrne & Kirlik, 2005). The theory found its application in researching human-computer interaction (see e.g., Fu & Pirolli, 2007). In education, ACT-R has been used to create so-called ‘cognitive tutors’ - artificial intelligence systems that serve to individualise and enhance learning. For example, the *Cognitive Tutor for Mathematics* is widely used in USA schools (Ritter et al., 2007). In the field of applied linguistics, ACT has been used to explain natural language processing including syntactic parsing (Lewis & Vasishth, 2005) and metaphor comprehension (Budiu & Anderson, 2002), while in the field of SLA, ACT theory has become known through works of Robert DeKeyser (2007; 2009) who relied on ACT to explain the cognitive processes underlying acquisition of a second/foreign language.

The CoP theory originated from Lave’s anthropological research into apprenticeship in communities (Lave & Wenger, 1991). This theory has been enthusiastically embraced to encourage learning both within organizational and educational settings. In academia, the notion of CoP has transformed the view on learning, shifting the focus from acquisition to participation (Hughes et al., 2007). Outside of academia, CoP has found practical applications in business, government, education, and organizational design (Wenger-Trainer & Wenger-Trainer, 2015). The CoP concept was further developed through ethnographic research into various communities in the workplace, which makes it particularly pertinent to the present study which also deals with learning in a professional setting. In recent years, the CoP’s relevance has increased further with the development of online learning environments. CoP’s affordances for describing learning in virtual environments is another reason for drawing on the theory to discuss findings in this study which researched item-writing skill development as it happened in an online training course.

The two sections that follow outline the main precepts of the two theories: Section 2.3.1 explains the premises of the cognitive view on learning drawing mostly on concepts developed within the ACT theory and complementing them with several concepts from cognitive theories close to ACT (Hayes-Roth et al., 1981; Spillman & Kirsner, 2005); Section 2.3.2 introduces the social perspective of learning through the prism of the CoP theory.

2.3.1 Learning as an individual process: Cognitive view on learning

Cognitive scientists are primarily concerned with how human cognition operates while learning a new skill. The cognitive approach assumes that the acquisition of any complex cognitive skill happens according to the same set of general mechanisms inherent to the brain. The main principles of skill development from the cognitive perspective – primarily according to the ACT theory in its latest version ACT-R (Anderson, 1993) - are presented below.

2.3.1.1 Knowledge representation

ACT-R distinguishes between two types of knowledge – declarative and procedural: “declarative knowledge is factual knowledge that people can report and describe, whereas procedural knowledge is knowledge people can only manifest in their performance" (Anderson, 1993a, p.18). Pieces of declarative knowledge, which ACT-R calls *chunks*, are added to the declarative memory "a chunk at a time" (Anderson, 1993a, p.25). The chunks can then be combined into complex hierarchical structures. However, acquiring declarative knowledge does not in itself guarantee skill acquisition – a conversion of declarative knowledge into procedural should occur for the acquisition to happen. Procedural knowledge "must be compiled from declarative knowledge through practice" (Anderson, 1993a, p.21). In the process of knowledge compilation, *production rules* are formed, defined as "the basic units of skills" (Anderson, 1993e, p.286). Production rules can then be directly executed from the production memory thus avoiding the phase of knowledge interpretation which is necessary when only declarative knowledge is relied on. Production rules are specific to particular tasks, but "have variables to allow them to apply in more than one situation" (Anderson, 1993e, p.286).

Why do we need these two types of knowledge encoding? ACT-R explains that declarative encoding enables the rapid learning of new information and storing it in a flexible form that allows for many different applications. It can also be used for analytic and reasoning purposes. The drawback of declarative knowledge is that its application is slow because "each fact must be separately retrieved from memory and interpreted" (Neves & Anderson, 1981, p.60). Procedural encoding, on the other hand, represents knowledge "as something that can be directly executed and so needs no costly interpretation phase" (Neves & Anderson, 1981, p.61). However, procedural knowledge cannot be analysed and is inflexible since, once a production rule has been formed, it cannot be changed and can only be applied in the form it was learnt.

2.3.1.2 The process of cognitive skill acquisition

Skill acquisition starts with learning declarative knowledge relevant to the skill. The simplest type of declarative knowledge is direct step-by-step instructions, but rules, examples, and information gleaned through problem solving also constitute declarative knowledge. Using declarative knowledge to perform a task "is sufficient to generate the desired behaviour to at least some crude approximation" (Anderson, 1996, p.217). The disadvantage of relying on declarative knowledge, however, is that it has to be interpreted: "[t]he interpretive productions require that the declarative information be represented in working memory, and this can place a heavy burden on working-memory capacity. Many of the subjects' errors and much of their slowness seem attributable to working-memory errors" (Anderson, 1996, p.231).

With practice, declarative knowledge is converted into procedural via the process of *knowledge compilation* (or *proceduralisation*). During proceduralisation, production rules are formed which allow to bypass the retrieval and interpretation of declarative knowledge, thus freeing the working memory and making the performance faster and more accurate. According to ACT-R, the most common way to form production rules is through practice by analogy, that is by using examples:

> … the examples illustrate the solution of a similar problem and the problem solver analogically maps the solution of the example onto a solution for the current problem. With repeated practice, however, general rules develop and the specific example is no longer accessed (Anderson et al., 1997, p.932)

Learning by examples, which ACT-R calls *analogy compilation,* is supported with empirical evidence: it was observed that, initially, trainees focused on examples while performing a new task but, with practice, stopped looking at examples and no longer mentioned them in verbal protocols (see e.g., Blessing & Anderson, 1996). Earlier ACT research indicated that "a production rule can be created after a single example" (Anderson, 1993, p.87); however, later research demonstrated that this is not always the case: "there seems to be a gradual shift from example-based processing to rule-based processing. Perhaps, each trial gives subjects another opportunity to encode a rule. Or perhaps rule-based processing and example-based processing compete as alternative means" (Anderson & Fincham, 1994, p.1338). Overall, the speed of proceduralisation might differ greatly and is not the same for all learners and all types of production rule.

The process of skill acquisition does not end with the formation of production rules. New production rules are weak and require *tuning.* Tuning involves rule strengthening when, with repeated practice, "better rules are strengthened and poorer are weakened" (Anderson, 1996, p.241). Tuning also involves "an improvement in the choice of method for performing the task" (p.241). ACT-R posits that tuning "is largely a matter of trial-and-error exploration. With experience, the search becomes more selective and more likely to lead to rapid success" (Anderson, 1996, p.241).

Within the ACT theory, Hayes-Roth et al. (1981) described a two-step learning cycle of production rule formation and tuning that happens during formal training. In the first step, learners take in and interpret input from the instructor by relating it to their existing knowledge. Learners then "operationalize the advice by transforming the declarative knowledge into executable or procedural forms" (Hayes-Roth et a., 1981, p.232). Several problems might then occur: learners might misunderstand the advice; the performance plans they have developed might fail to work or might be difficult to execute. The awareness of the problems will trigger the second step of the cycle, when the learners "diagnose the problems in behaviour and refine the knowledge that underlies them" (Hayes-Roth et a., 1981, p.233). New knowledge might have to be learnt and new performance plans developed to refine the performance.

The ultimate goal of tuning is production rule *automatization*. However, while the initial rule strengthening might take only several successful attempts, automatization requires a long period of deliberate practice. Automatization is often associated with expert performance, and the '10-year rule' has been observed for many cognitive skills, whereby ten years of deliberate practice are necessary to reach expert status in a particular domain (Proctor & Dutta, 1995).

2.3.1.3 Learning trajectories

ACT-R acknowledges the power law of practice which postulates that the speed and accuracy of performance improves through practice, but the speed of improvement is uneven: it is fast initially but slows down with time, finally reaching the stage when no further improvement is observed (*Figure* 2-1).

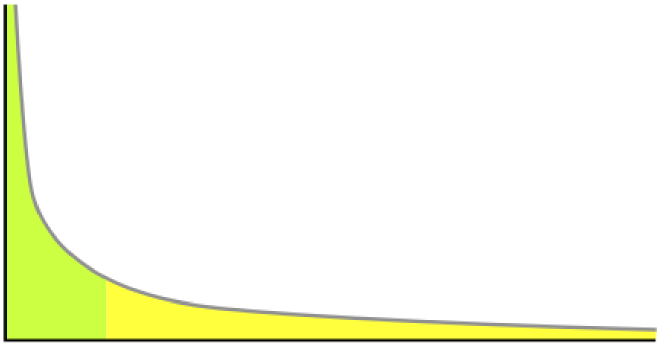


Figure 2-1. Example of power law. X axis represents learning time, Y axis represents performance time. Kranen, H. (2006). [Long tail] [graph]. Wikipedia. https://en.wikipedia.org/wiki/Power_law#/media/File:Long_tail.svg

Although the power law is well-established in psychology, its universality has been repeatedly challenged. Early into the development of the skill acquisition theory, Fitts and Posner (1967) observed that "performance does not inevitably improve with practice" (p.18) and learning curves for different learners might differ: some learners will demonstrate fast improvement with exponential function as a better fit, while for some learners the improvement will be much slower or even stalled. Moreover, for complex skills, different skill components will develop at different speeds, which is in line with the ACT-R view on skill acquisition as "a process of continual refinement … of a rather complex system of interactions" (Anderson, 1996, p.255). ACT-R posits that not all production rules of which a skill is comprised are formed and then tuned at the same time, which often results in a situation where "part of a task can be performed interpretively [i.e. using declarative knowledge] while another part is performed compiled [i.e. using production rules]" (Anderson, 1996, p.255).

Speelman and Kirsner (2005) wrote that "learning rate is affected by the relative amounts of practice of particular task components and also the relative number of processing steps involved in these components" (p.80). Further, Speelman and Kirsner (2005) believed that a complex skill might contain components from previously learnt skills, while some components might have to be learnt anew. Consequently, the performance on the previously learnt components will be better than on the new ones, while the previously learnt components will also have less room for improvement. At the same time, a "change in task conditions" (Speelman & Kirsner, 2005, p.100), when a previously learnt component is integrated into a new skill set, will lead to a decline in performance during the transfer process.

Another phenomenon, called the *learning curve plateau* (Figure *2-2*), has been repeatedly observed, in particular when different components of a complex skill are acquired at different rates. The plateau effect was first discovered through research into telegraphy skills, when Bryan and Harter (1899) noticed periods of no change in some telegraphers' performance, while for other telegraphers the performance temporarily declined.

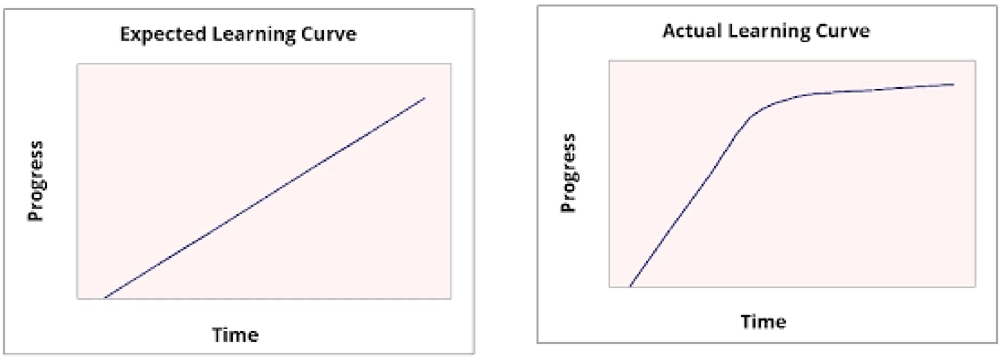
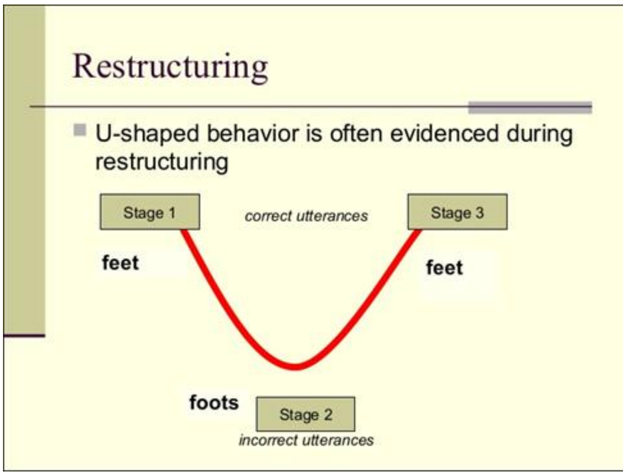


Figure 2-2. Expected vs actual learning curve: Plateau effect. van Vliet, N. (2015). [Learning curves] [graph]. Smart Language Learner. https://www.smartlanguagelearner.com/wp-content/uploads/2015/09/expected-learning-curve1.jpg

Furthermore, since that early research, *U-shaped* or *non-monotonic* skill development has been observed in many areas of learning including SLA (McLaughlin, 1990), child development

(Strauss & Stavy, 1982), and medical education (Lesgold et al., 1988). McLaughlin (1990) named the U-shaped learning curve in SLA *restructuring* (see an example in *Figure 2-3*):

> … practice can lead to improvement in performance as sub-skills become automated, but it is also possible for increased practice to create conditions for restructuring, with attendant decrements in performance as learners reorganize their internal representational framework. In the second case, performance may follow a U-shaped curve, declining as more complex internal representations replace less complex ones, and increasing again as skill becomes expertise. (p.113)



Restructuring

- U-shaped behavior is often evidenced during restructuring

Figure 2-3. An example of restructuring. Rosete, R. (2013). [Psychology of language] [PowerPoint slide]. Slideshare. https://image.slidesharecdn.com/psychologyoflanguage-130801145650-phpapp02/95/psychology-of-language-15-638.jpg?cb=1375369075

In language testing, the U-shaped learning curve has recently been reported in rater training. Yan and Park (2019) observed U-shaped fluctuations in rater performance while the newly trained raters were exploring different rating strategies and forming “their own interpretation and operationalization of the rating scale” (p.24).

Strauss and Stavy (1982) suggested several reasons for U-shaped skill acquisition:

- the learner has two production rules for the same task – a familiar but inadequate one and a new but ‘untrusted’ one – and oscillates between the two;
- the learner uses the production rule for a familiar task in performing a different new task for which the rule is inadequate;
- the learner has acquired all necessary components for a complex skill but is not yet able to co-ordinate these components.

More than one explanation can apply for a U-turn in a learner's performance. Moreover, a U-turn and/or a plateau, although they occur for some learners, do not happen to everyone: "[t]he current view is that plateaus do not represent a necessary stage of learning" (Proctor & Dutta, 1995, p.8).

2.3.1.4 Transfer of training

How mastering one skill can help in acquiring a different skill has been the focus of research for a long time. Thorndike and Woodworth (1901) put forward *a theory of identical elements*, whereby transfer will happen if two tasks share common elements. ACT-R suggests that production rules are behind the transfer of training and describes the transfer in a similar way as Thorndike and Woodworth (1901) did: the larger the overlap in production rules between the two tasks, the more extensive the transfer. ACT-R predicts the possibility of three types of transfer: (1) *positive transfer* which happens "to the extent that the two skills involve the same productions" (Anderson, 1987, p.198); (2) *negative transfer* when "the productions optimal for one skill might transfer to another skill where they are no longer optimal" (Anderson & Singley, 1993, p.191); and (3) *zero transfer*, when two skills have no productions in common, so "learning of the second would proceed as if the first had not been learnt" (Anderson & Singley, 1993, p.191).

Speelman and Kirsner (2005) further developed the idea of transfer by distinguishing two types of production rules – general and specific. A general production rule, once acquired, can be applied to perform similar but different tasks, while specific production rules can only be used to perform the same type of task. Speelman and Kirsner (2005) also observed that the act of transfer has the initial effect of slowing down the performance: "rapid rates of improvement ... co-occur with low levels of transfer... It is the child or adult with no practice on related tasks that should show rapid early progress" (p.16).

2.3.1.5 The role of training in learning

ACT-R's training principles are largely based on research into 'intelligent tutoring systems' used to teach programming languages, advanced maths, and problem-solving skills (Anderson & Corbett, 1993). The most important of the principles is that "the skill itself should be modelled as a set of production rules" (p.237). This approach assumes that the production rules underlying a skill are known to the trainers, therefore the first step in training design is "to come up with a set of production rules that represent the skill we want the student to master" (p.237). ACT-R warns that the task is not easy because the production rules must "capture the complexity of the domain" (p.237). ACT-R concedes that there might be multiple efficient ways of performing a task and multiple production systems; however, the training effect is greater if students are taught "more powerful ways to solve problems" (p.238).

Additional recommendations for cognitive skill training can be found in works of cognitive psychologists close to ACT-R. For example, Fitts and Posner (1967) believed that skill acquisition "rarely lives up to the potential of the subject" (p.26) and "it is necessary to maintain the subject's motivation, to provide him with knowledge of results, and to take into account extraneous limitations" (p.18). One of the reasons for inferior learning might be information overload which can lead to learners filtering out parts of the information necessary for successful performance. Therefore, training should allow for task repetition followed by feedback which "serves as a powerful reinforcer in the learning of skills" (p.28).

Speelman and Kirsner (2005) argued that any skill training should aim for transfer by prioritising the development of general production rules which can be applied to a greater variety of individual tasks. The development of general production rules can be encouraged with task variety because the variety leads to task comparison and "abstraction of features that are common to many items" (p.74). Proctor and Dutta (1995) described a *contextual interference effect* when the order of task practice influences skill retention. If a task involves several forms, the training can be done either in blocked order (one form of the task is practised several times before the training moves to a different form) or random order. Blocked order results in faster learning as measured during the training, but random order leads to better retention and to the skills being more generalizable.

This section outlined the main premises of a major cognitive learning theory ACT-R: differences between declarative and procedural knowledge, the process of skill acquisition, the power law of practice and deviations from it – the learning curve plateau and the U-turn, the possibility of training transfer, and the role training plays in skill acquisition. The section that follows will discuss the main premises of the social view on learning as embodied in the theory of Communities of Practice.

2.3.2 Learning as a social process: Situated view on learning

The Communities of Practice (CoP) theory, similarly to other social learning theories, "has its roots in attempts to develop accounts of the social nature of human learning" (Wenger, 2010, p.179) by adopting a perspective fundamentally different from – but not incompatible with – the one of ACT-R. In fact, the two theories are complementary because they deal with different dimensions of a multidimensional phenomenon. CoP has its unique set of assumptions about knowledge, learning, and the learner, which are outlined below.

2.3.2.1 Communities of practice

Wenger (1998) wrote that CoPs are everywhere – every person can expect to belong to multiple CoPs throughout his/her life. His most recent definition of CoPs is as "groups of people who share a concern or a passion for something they do and learn to do it better as they interact regularly" (Wenger-Trayner & Wenger-Trayner, 2015, p.1). A CoP does not have to be a well-defined formal group, though, and belonging to a CoP is not about socially-visible boundaries but about "participation in an activity system about which participants share understandings concerning what they are doing and what it means" (Lave & Wenger, 1991, p.98).

CoPs have three dimensions:

- **Domain:** a sphere of activity that the community is engaged in. Snyder and Wenger (2010) emphasize that "passion for the domain" (p.110) is crucial for the feeling of belonging to the community;
- **Community:** includes the community itself and the relationships between its members. The quality of relationships defines the strength of the community;
- **Practice:** a community exists by being active in and developing the knowledge of a domain-relevant practice. The practice includes a repertoire of frameworks, methods, tools, and activities.

CoPs are "the social 'containers' of the competencies" (Wenger, 2000, p.229). A competence, in the CoP sense, includes three elements:

- **Joint enterprise**: the sense of belonging to a community and understanding how it works;
- **Mutual engagement**: the act of engaging with the community and being perceived as its trusted member. Importantly, it is not geographical proximity that defines engagement but the level of meaningful interaction among its members.
- **Shared repertoire**: access to the language, frameworks, routines, and tools that the community uses.

In the CoP theory, practice is not simply an activity a community is engaged in but “a form of belonging. Such participation shapes […] who we are and how we interpret what we do” (Wenger, 2012, p.292). Practice has two components:

- **Participation:** “both action and connection” (Wenger, 1998, p.57); partaking in the community’s activities and building/maintaining relations with others within the community;
- **Reification:** “giving form to our experience by producing objects that congeal this experience into ‘thingness’” (Wenger, 1998, p.58). Reification might involve producing tools, instructions, a set of terminology, or informal stories related to the practice the community is engaged in.

Wenger (1998) believed that, for the successful functioning of a CoP, participation and reification should be in balance. If participation prevails – that is the community’s activity is based on unwritten rules and non-standard practices – it might be difficult to co-ordinate members’ activity. On the other hand, if everything is reified, there is “little opportunity for shared experience and interactive negotiation” (Wenger, 1998, p.65).

2.3.2.2 Learning as legitimate peripheral participation

Wenger’s social view of learning is based on four premises:

1. People are social beings;
2. *Knowledge* is competence in something that matters;
3. *Learning* means actively pursuing that competence by engaging with the relevant community;
4. Successful learning results in *meaning* defined as “our ability to experience the world and our engagement with it” (Wenger, 2012, p.291) in a meaningful way.

Learning as a social activity situated within a CoP can be called *legitimate peripheral participation* (LPP) - the process whereby “newcomers become part of a community of practice” (Lave & Wenger, 1991, p.29) by gradually increasing their participation in the CoP’s activities while learning the knowledge and skills that characterise the CoP’s practice. LPP “moves in a centripetal direction, motivated … by the growing use value of participation, and by newcomers’ desires to become full practitioners” (Lave & Wenger, 1991, p.122).

In CoP’s view, it is a community and not books that holds knowledge, therefore traditional formal learning whereby students are removed from the community and put into a classroom where teachers feed them with knowledge derived from books is not effective. Instead, it is “engaging students in meaningful practices … providing access to resources …. opening their horizons … involving them in actions, discussions, and reflections” (Wenger, 2018, p.225) that can result in useful learning. It is only through LPP that learners’ identities become engaged, while “interaction with acknowledged adept practitioners makes learning legitimate and of value … learners know that there is a field for the mature practice of what they are learning to do” (Lave & Wenger, 1991, p.110).

2.3.2.3 Cultivating CoPs to promote learning

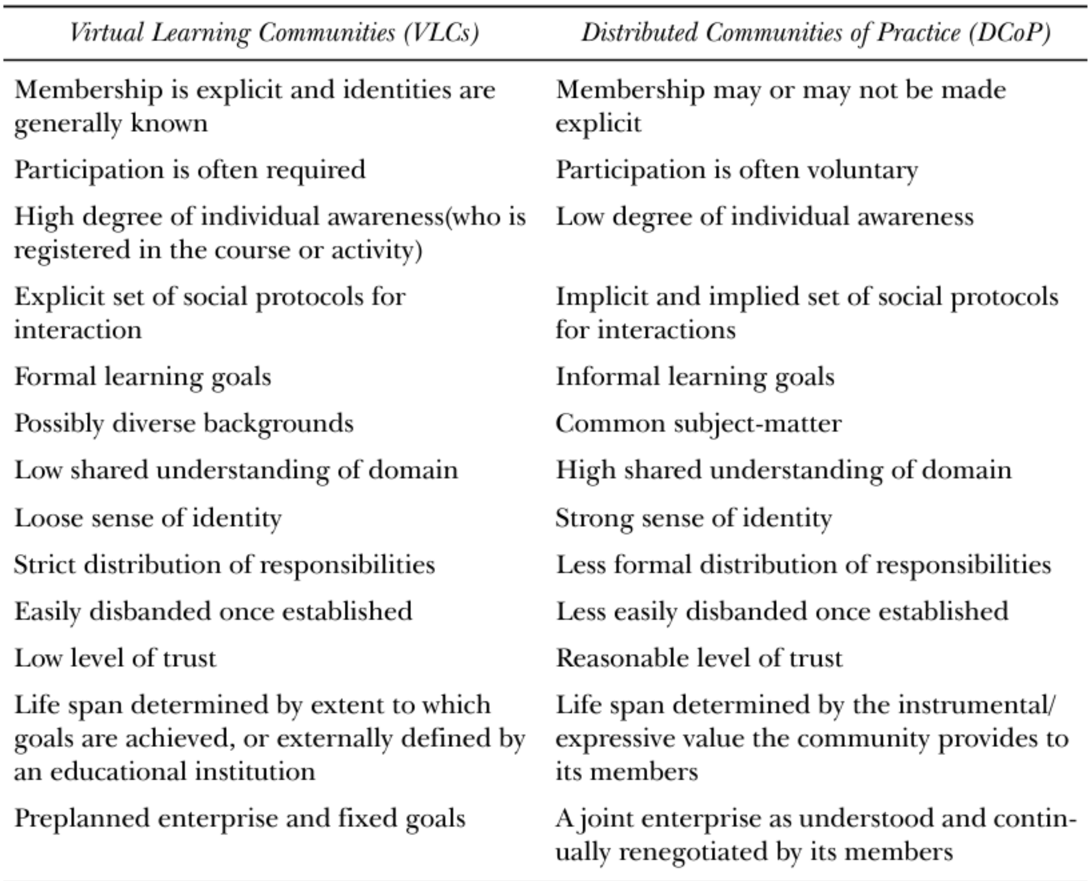
Wenger at al. (2002) proposed seven general principles of cultivating CoPs that are applicable both to working and educational environments:

1. Design for evolution.
2. Open a dialogue between inside and outside perspectives.
3. Invite different levels of participation.
4. Develop both public and private community spaces.
5. Focus on value (of the community and its members).
6. Combine familiarity and excitement.
7. Create a rhythm for the community.

(Wenger et al., 2002, p.50)

The original CoP concept as developed by Wenger (1998) does not make a hard distinction between a CoP and a learning community, positing that a CoP is the primary place where learning should happen. Hoadley (2012), although using the term CoP for both types of community, warned that “we must be careful to distinguish between a community of practice as a phenomenon (naturally occurring or otherwise), versus an intended or designed learning environment” (p.295). However, as the original CoP concept was undergoing changes, the two types of community came to be seen as different. Other refinements to the theory were also proposed; for example, the concept of a distributed CoP (DCoP) was developed, whereby CoP members are distributed in space (see e.g., Schwier & Daniel, 2008). Wenger (1998) himself, however, did not originally differentiate between a CoP and a DCoP, believing that it is not geographical proximity but "dense relations of mutual engagement" (p.75) that define community members' interaction. The notion of a DCoP is primarily used to describe learning in virtual environments.

Differences between a learning community and a CoP, as well as between a virtual learning community (VLC) and a DCoP have been scrutinised (see e.g., *Figure 2-4*). The main difference is that "the learning community is an artificial construct created by the teacher with a didactic goal" (Bos-Ciussi et al., 2008, p.303). In a learning community, there is a tension between an obligation (imposed by the teacher) and a necessity (emerging from students' needs) – for a learning community to become a CoP, the necessity should prevail over the obligation.



Virtual Learning Communities (VLCs)	*Distributed Communities of Practice (DCoP)*
Membership is explicit and identities are generally known	Membership may or may not be made explicit
Participation is often required	Participation is often voluntary
High degree of individual awareness(who is registered in the course or activity)	Low degree of individual awareness
Explicit set of social protocols for interaction	Implicit and implied set of social protocols for interactions
Formal learning goals	Informal learning goals
Possibly diverse backgrounds	Common subject-matter
Low shared understanding of domain	High shared understanding of domain
Loose sense of identity	Strong sense of identity
Strict distribution of responsibilities	Less formal distribution of responsibilities
Easily disbanded once established	Less easily disbanded once established
Low level of trust	Reasonable level of trust
Life span determined by extent to which goals are achieved, or externally defined by an educational institution	Life span determined by the instrumental/ expressive value the community provides to its members
Preplanned enterprise and fixed goals	A joint enterprise as understood and continually renegotiated by its members

Figure 2-4. Key features of VLCs and DCoPs (Schwier & Daniel, 2008, p.350)

Bos-Cuissi et al. (2008) provided some recommendations for teachers on how to cultivate a CoP in virtual learning environments: the teacher should stay in the background, create learning content that encourages students to interact, and "set up strict rules in order to encourage exchanges to emerge" (p.303). However, the latter can also be problematic because forcing students to interact might once again lead to teacher-domination thus destroying the sense of community. Hibbert (2008) acknowledged this threat by saying that "a sense of community cannot be mandated or forced but conditions can and must be created to promote its development" (p.143). She argued that teachers can cultivate DCoPs by modelling online presence, activating meaningful participation, promoting the development of relationships, building community-oriented – not simply task-oriented – discussions, and creating informal space within the course where students can chat informally (p.143). Overall, as the CoP concept gained popularity as a means of promoting learning and, in particular, online learning, a large number of publications appeared reporting on empirical studies and offering practical advice on how to cultivate CoPs in learning communities – see, for example, edited volumes by Kimble, Hildreth, and Bourdon (2008), and by Land and Jonassen (2012).

This section outlined the main premises of a popular social learning theory (CoP), i.e. dimensions and elements of a CoP, legitimate peripheral participation as the main learning route within a CoP, and the principles of cultivating a CoP to promote learning.

2.4 Chapter summary and research questions

As argued in this chapter, item quality makes a significant contribution to test validity, thus making item writing critically important for assessment. However, operational aspects of item writing are still generally overlooked in research and publications. In particular, while assessment specialists have long been advocating the necessity for item-writing training in order to ensure item quality, little research has so far been done - neither in language testing nor in educational measurement in general. The language testing organisations that have reported on their item-writing training procedures did so without providing any critical reflection or evaluation of the training's effectiveness. Those few empirical studies that have so far investigated item-writing training effectiveness – none of which are from the field of language testing – produced unconvincing results, partly due to multiple methodological flaws and partly due to the impossibility of drawing practical implications from the studies. Moreover, the very process of writing test items is still little understood. Few attempts have so far been made to empirically investigate it, none of which has resulted in a comprehensive description of item-writing expertise. The studies have also acquired a static view on item-writing expertise with no research looking into how item-writing skills develop as a result of training.

To address these research gaps, the present study aims to empirically explore item-writing skills and their development as it happened during an online induction item-writing training course. By looking into changes in the quality of items produced by course participants before and after the training, as well as by analysing participants' accounts of their item-writing experiences and of their perceptions of training effectiveness, the study strives to gain insights into item-writing skill development and, in particular, into the role of induction training in it. Two influential learning theories will be drawn upon to interpret the insights gained in this research project: the ACT-R theory for the cognitive side of item-writing skill acquisition, and the CoP theory for its social dimension.

The following research questions have been formulated:

RQ1: How did the quality of items produced by novice item writers change from before to after an online item-writing training course?

RQ2: How did the participants' item-writing skills develop through the training, as perceived by the participants in interviews?

RQ3: What role did the participants perceive the training played in their item-writing skill development?

The research questions are operationalised through a mixed-methods research design. The Methodology Chapter that follows describes the study's research design and provides a detailed overview of its operationalisation in the present research project.

Chapter 3 Methodology

3.1 Introduction

In order to provide answers to the research questions, a mixed-methods approach was used consisting of a Pretest-Posttest study and a Course Feedback study, involving both quantitative and qualitative data collection and analyses. Expert judgement, interview, and feedback questionnaire data were triangulated to ensure methodological validity as well as to provide reliable and rich answers to the research questions.

This chapter provides an account of each research method used in the study. Section 3.2 contains an overview of the online item-writing training course, of which this study aims to explore the effects. Section 3.3 outlines the overall research design and describes the procedures used to ensure that the study complied with research ethics in social sciences. Section 3.4 gives a brief overview of the pilot study that was conducted to trial research instruments and procedures. The main study is detailed in Section 3.5, including the Pretest-Posttest study (3.5.1) and the Course Feedback study (3.5.2). Finally, Section 3.6 summarizes the contents of this chapter.

3.2 Online item-writing training course overview

This research project was carried out during two cohorts of an online item-writing training provided to British Council China employees. The employees were based in China and were encouraged to develop their language assessment-related skills to contribute to various assessment projects the organisation was involved in in the East-Asia region. The only feasible mode of course delivery was online because the employees were based in four different locations across China and had to travel regularly for work. Three item-writing training cohorts were taught: Cohort 1 was taught between October 2016 and February 2017 and constituted an induction cohort during which the course concept and the first version of course materials were developed. Cohort 2, which formed the basis of the pilot study, was taught between May-August 2017 and had only 10 participants. The main study described in this thesis was conducted with Cohort 3, taught between December 2017 and May 2018. It was the largest of the three cohorts, with 25 participants who completed all course requirements.

Importantly, the course evolved with each run, with the course structure tweaked, some new materials created, and some old materials replaced or revised based on previous runs, tutor observations, and participant feedback. Therefore, the data collected during Cohort 2 and Cohort 3 are not fully comparable, and the pilot study with Cohort 2 was used to trial the research instruments and procedures rather than for comparative analysis of findings.

My involvement with the course requires some explanation. Just prior to my doctoral studies at Lancaster University, I was working as a British Council China employee and was approached by the management with an offer to create and run an online item-writing training programme. I felt I had the necessary skills and qualifications – an MA in Language Testing, several years of relevant item-writing experience, as well as e-moderator qualifications – so I embraced the opportunity of what promised to be a very interesting and developmental experience. I had to design the course from scratch, including the creation of all materials (see Section 1.2.2 for information on the theoretical framework that underlies the development of item specifications used in the course, and Section 1.2.3 for an outline of the pedagogical principles that informed the course design). I also acted as the sole course tutor for Cohorts 1 and 2, while I had a co-tutor during Cohort 3. While teaching Cohort 1, I realized that item-writing training is an area that, on one hand, interests me both as a trainer and an assessment researcher, while on the other hand, remains largely under-researched (see the Literature Review Chapter).

The online item-writing course I developed was a three-month training programme consisting of six modules (the exact number of modules changed from cohort to cohort, but the general structure remained very similar). A four-to-five hour time commitment was expected from the participants each week. The course aimed to give participants theoretical knowledge and to develop practical skills in writing a broad range of language test items (see Section 1.2.1 for information on how the scope of the training was determined). The course included theoretical input, group discussion activities, and item-writing practice. Course modules covered the following topics: introduction to item-writing and the CEFR, producing grammar and vocabulary items, and producing items for the four language skills (speaking, writing, reading, and listening). Each module ran over two weeks: first, participants were introduced to the item-writing topic under focus, learnt about item-writing techniques for specific types of items and/or specific language areas/skills, and discussed successful/problematic items and their characteristics. In week 2, participants wrote their own items according to specifications, peer-reviewed them in small groups, and then submitted the revised versions to the course tutor(s) for individual feedback. The course syllabus, including module topics, activities, materials, modes of interaction and types of feedback, can be found in *Appendix 1.* To illustrate, the structure of Module 6 is discussed in the next paragraph.

Module 6 was dedicated to writing listening test tasks. In week 1, participants studied two PowerPoint voice-over presentations recorded by the course tutor. The first presentation covered the construct of listening assessment, higher- and lower-level listening processes (Field, 2013), what makes listening difficult (Green, 2017), and differences between spoken and written language including phonological, lexical, grammatical, and discursive characteristics of spoken texts (Carter & McCarthy, 2007; Wagner, 2016). In another recorded presentation, participants were introduced to practical techniques for developing listening input texts. These included the 'textmapping' techniques of exploiting genuine sound files (Green, 2017), semi-scripting (Buck, 2001), and introducing spoken language characteristics into scripted texts (cf. Wagner, 2018). Following the presentations, participants completed two practical tasks: 1) reflecting on the authenticity of the listening input text each participant had developed as part of the pre-course assignment, revising the text to make it more authentic-sounding, and posting the revised text in their discussion groups for peer-feedback; 2) 'textmapping' a genuine sound file provided by the tutors and, through a group discussion activity, arriving at a consensus about the file's gist to be targeted in items. During week 2, participants were first introduced to principles and techniques of producing listening test items in another recorded presentation. Each participant then developed three listening tasks, including texts and items, according to a set of specifications provided by the tutors. Participants discussed their tasks in groups with an opportunity to revise them before submission to the course tutors, who then provided detailed individual feedback. As a follow-up, the tutors posted a feedback summary to the course platform discussing common problems and offering further advice.

The asynchronous online delivery happened through *Edmodo*, a free Learning Management System (LMS). Within *Edmodo*, all training materials were uploaded to the course library, while module instructions and feedback summaries were posted on the course page. For each module, participants were divided into small groups of four-six people to discuss tasks and give feedback on each other's items. *Wechat* (the Chinese equivalent of WhatsApp) was used to facilitate online group discussions. Course assignments were submitted via email.

Participants took the course on a voluntary basis. All employees who expressed interest in the next cohort did a pre-training assignment consisting of several item-writing tasks. They submitted the assignment within one week and, upon submission, were considered enrolled.

The very fact of submission and not the quality of their items qualified participants for the course. To receive a course completion certificate, participants had to submit a post-training assignment. The assignment included tasks in all areas/skills covered during the course. All assignments were graded by the course tutors against detailed checklists. Grades were converted into percentages and included in course certificate transcripts. The post-training assignment results were also communicated with the British Council China management to inform participant recommendations for future item-writing work.

3.3 Overall research design

The three research questions this study aims to answer were operationalized through a mixed-methods research design (*Figure 3-1*), involving the Pretest-Posttest and Course Feedback studies which drew on both quantitative and qualitative data obtained using three different data collection methods. More specifically, to address RQ1 (*"How did the quality of items produced by novice item writers change from before to after an online item-writing training course?"),* expert judgements were obtained on the quality of items participants produced for the pre-training and post-training assignments. The quantitative data from these item evaluations was analysed using statistical methods. To address RQ2 (*"How did the participants' item-writing skill develop following the training, as perceived by the participants in interviews?"),* interviews were conducted with participants upon completion of their pre- and post-training assignments. The interview transcripts were coded and analysed using the Grounded Theory approach. Finally, to address RQ3 (*"What role did the participants perceive the training played in their item-writing skill development?"),* feedback questionnaires were administered to participants throughout the course. The quantitative and qualitative data from the questionnaires was combined with the findings from that part of the post-training interviews where participants provided feedback on the training. Descriptive statistics were calculated for the quantitative data, while the qualitative data was coded and analysed thematically. In the following paragraphs, the study's research design is detailed and justified with reference to the research methodology literature.

Mixed-methods research designs have been defined as "research in which the investigator collects and analyzes data, integrates the findings, and draws inferences using both qualitative and quantitative approaches or methods in a single study or a program of inquiry" (Tashakkori & Creswell, 2007, p.4). Mixed methods approaches have a number of advantages: they can provide answers to both confirmatory and exploratory questions thus widening the research scope; they provide stronger inferences through data triangulation; they allow the researcher to deal with divergent findings in situations where “quantitative and qualitative components lead to totally different (or contradictory) conclusions” (Teddlie & Tashakkori, 2009, p.35).

This research project can be classed *as multistrand* (Teddlie & Tashakkori, 2009) because it incorporates two studies with each study including two strands, all integrated at the final stage of analysis (*Figure 3-1*). The Pretest-Posttest study consisted of two data collection phases: one before the item-writing course (a pre-training assignment), and one after the item-writing course (a post-training assignment). The data for the Course Feedback study was mostly collected while the course was on-going, except for the interview data which was collected after the post-training assignment. Data analyses began after all data had been collected and was both quantitative and qualitative in nature: item quality was evaluated quantitatively against a rating scale, and item ratings were then analysed using statistical methods; transcripts of interviews were analysed qualitatively, while feedback questionnaire data was analysed both quantitatively and qualitatively. Different types of data were first analysed independently and then brought together for the final stage of analysis, where the findings were combined and triangulated.

3.3.1 Ethics

As this study involved human participants, ethics approval was sought from and granted by the FASS-LUMS Research Ethics Committee at Lancaster University. Information Sheets (*Appendix 2*), and Consent Forms (*Appendix 3*) were developed for the two groups of participants in this study – item-writing trainees and expert judges. The information sheets explained the aims and nature of the study, and informed prospective participants of their right to decline participation or withdraw from the study. Written consent was obtained from all participants.

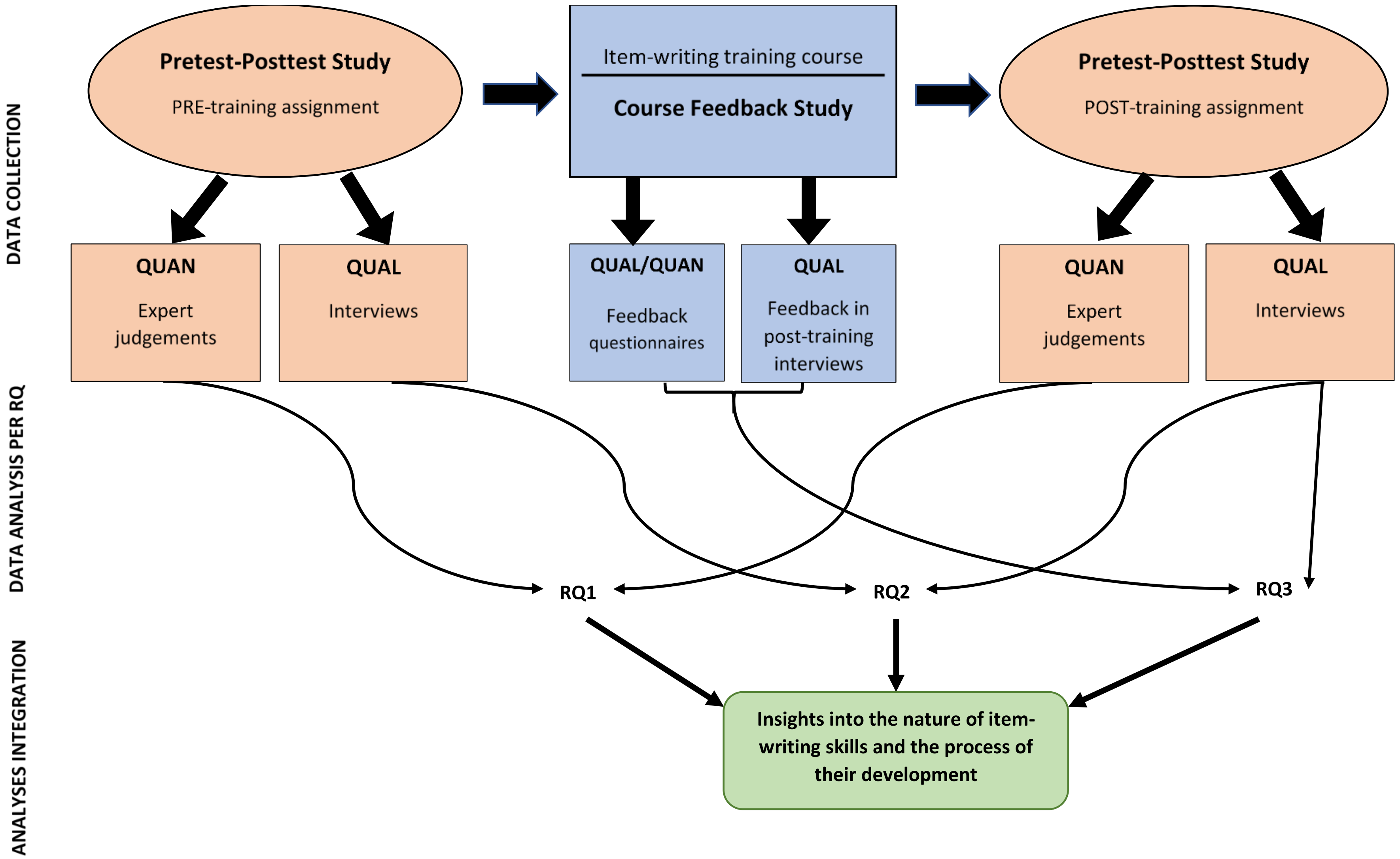


Figure 3-1. Overall research design

3.4 Pilot study

The pilot study was conducted between May-August 2017 with Cohort 2. It had 10 participants, three of whom were female and seven male, age range 28-60 (M=42). All participants were British Council China employees. All were native speakers of English (American, Australian, British, and South African) educated to a minimum of BA level with six also having an MA-level degree. All participants were qualified ESL/EFL teachers, with 5-36 years teaching experience (M=13). None had written test items for a professional exam board prior to the training, but all had some classroom assessment experience and five participants reported having been involved in creating language tests for classroom use or university entry.

The pilot study served two important purposes : a) to trial research instruments and procedures for the main study, and b) the item evaluation scores from the pilot, together with the scores from the main study, were used to calculate Intraclass Correlation Coefficients (ICC) to establish expert judges' agreement. The former use is outlined below, while the latter is explained in Section 3.5.1.4.3.

All data collection instruments were trialled during the pilot study and subsequently revised:

1. A *background questionnaire* was designed to collect information about participants. Based on pilot study responses, the background questionnaire was slightly edited, with some questions reworded to make them clearer, but otherwise the questions remained the same (see Section 3.5.1.3 for information about the Background questionnaire content).

2. An *item-writing assignment* was administered to participants both before and after the training. After piloting, the assignment underwent minimal changes: slightly altering the layout and font colour of several sentences in the instructions. Section 3.5.1.2 provides detailed information regarding the content and the administration of the assignment.

3. *An item evaluation scale* was developed to evaluate the quality of items produced by participants. The scale underwent substantial revision after piloting, as described in Section 3.5.1.4.

4. An *interview schedule* was drawn up to interview participants upon completion of the item-writing assignments. Participants' pilot responses were analysed for clarity and helpfulness of questions, the questions' ability to elicit relevant information was

reviewed. The interview schedule underwent some changes after piloting, as described in Section 3.5.1.5.

5. *Four feedback questionnaires* were designed to obtain participants' opinions about the item-writing training course. The pilot responses were analysed for clarity and usefulness of questions, and the questionnaires were subsequently revised (see Section 3.5.2.1 for more details).

3.5 Main study

The two sections that follow provide a detailed account of the two studies of which the research project's main study is comprised. They describe the data collection and analyses methods for each type of data involved. More specifically, Section 3.5.1 focusses on the Pretest-Posttest study consisting of two strands, quantitative expert judgements and qualitative interviews, while Section 3.5.2 provides an account of the data collection and data analyses for the Course Feedback study.

3.5.1 Main study 1: Pretest-Posttest

The pretest-posttest study consisted of two phases – one before and one after the training, both involving quantitative item evaluations and qualitative interviews. This section first provides a brief overview of quasi-experimental research designs (3.5.1.1), then describes the item-writing assignment which formed the basis for data collection (3.5.1.2), and the item-writing trainees (3.5.1.3). Sections 3.5.1.4 and 3.5.1.5 describe the methods of data collection and data analyses for the expert judgements and the interviews, respectively.

3.5.1.1 Quasi-experimental research design

An experimental study "involves an experiment in which data are collected in two or more conditions that are identical in all aspects but one" (Chow, 2010, p.447). A "*classic true experiment*" (Cohen et al., 2011, p.315) comprises an intervention and two groups of participants – an experimental and a control group – with participants randomly assigned to each group. Although such experiments are widely used in laboratory settings, real-life environments often make true experiments impossible due to practical or ethical considerations, in which case a quasi-experiment can be considered[3] as an alternative. *Quasi-experimental* designs are used when random allocation of participants to the experimental and the control group is not possible, which is often the case in educational settings (Cohen et al., 2011). Also, in many educational studies, for example for course evaluation purposes, having a control group is not feasible (Lynch, 2003). In such cases, a *quasi-experimental design* with a single experimental group is employed, and the most typical research design in this case is a *one-group pretest-posttest design* (Baldwin & Berkeljon, 2010) which can be schematically represented as "O1 – X – O2" where O1 is a pre-test measure, X is an intervention, and O2 is a post-test measure (p. 1173).

The present study adopted such a quasi-experimental *one-group pretest-posttest* research design. A group of self-selected participants (see Section 3.5.1.3) willing to acquire item-writing skills volunteered to take an item-writing training course. As a *pre-test* measure, they completed a pre-training assignment consisting of three item-writing tasks. They then took a three-month item-writing course *(intervention)* and, following the training, they completed a post-training assignment *(post-test)* that contained tasks identical to the pre-training ones (see Section 3.5.1.2).

A drawback of quasi-experimental designs is their weak internal validity: because there is no control group for comparison as well as no random allocation of participants to the group, "alternative explanations are difficult to rule out" (Baldwin & Berkeljon, 2010, p.1171). If we take a vocabulary learning experiment as an example, it would be difficult to determine, having only pre/post test results from one experimental group, whether the learning occurred as a result of the intervention or due to other variables, such as incidental learning that happened outside the experiment, natural subject maturation (which is often the case in longitudinal studies with young learners), or a 'testing practice effect' (Glass, 1965) whereby "persons tend to increase their scores on second and subsequent administrations of a test because of familiarity with the format of the test, recalling answers to items, etc." (p.83).

Despite, and in full realization of, this disadvantage, a quasi-experimental design was argued to be the only suitable and feasible one for the present study. Wheelan (2013) discusses "cases where a controlled experiment is impractical or immoral" (p.240), and the present study carries both characteristics. From a practical point of view, it would be difficult to find 25

[3] Sometimes two different terms – 'quasi-experimental' and 'pre-experimental' – are used, see for example the *Encyclopedia of Research Design* (2010). However, both terms refer to the same approach, i.e. an intervention without random allocation and, sometimes, in the absence of a control group.

control group participants with a background similar to the experimental group: as it was an internal training, they would have to be employees of the organization, educated to a minimum of BA level, having an EFL/ESL background, with similar teaching and examining experience. Even if such participants could be found, they would have to be offered a reasonable incentive (for which I did not have the means) to devote hours of their time to writing a large number of test items for the pretest-posttest assignments. However, even if funds were available, the research conditions would become either unequal or unethical: while experimental group participants would take part in the study motivated in mastering a new skill and become professional item writers and thus would put full effort into their work, the control group would be motivated through obtaining a financial reward on completion of the task, irrespective of the quality of their work, so they might perform the item-writing task in a perfunctory manner. On the other hand, if it were possible, at any one time, to find 50 participants who would have equal desire in becoming item writers, and then allocate them to either the experimental or the control group, the study would become questionable from an ethical perspective: while the experimental group participants would receive the benefit of the training they had desired (with potential job prospects), the equally motivated control group would be deprived of such an opportunity, and instead would be requested to write two sets of test items without receiving any feedback or information about their work.

As for the quasi-experimental method's risks mentioned earlier, it is unlikely in the present study that learning would occur outside the experiment. Item-writing skills are unlikely to be picked-up or developed elsewhere other than during item-writing training or item-writing practice, because none of the participants had access to other forms of item-writing experience at the time of the study, for example by writing items for classroom tests. As for a 'testing practice effect', this possibility cannot be excluded, since the tasks for both pretest and posttest assignments were identical to allow the resulting items to be randomized for expert judgement evaluation later. Moreover, as part of the training, the participants were provided with feedback on their pre-training items, so if they chose to pay attention, they had the opportunity to learn from the pre-test to improve on their post-test items. Therefore, because of the nature of this study, where the pre-test is part of the learning process and not a laboratory experiment, the said disadvantage cannot be avoided; this is a situation typical of many empirical studies conducted in educational settings (Lynch, 2003). However, the present study has a wider aim than just comparing pre- and post-test results for evidence of learning gains. It aims to look at the trajectory of item-writing skill development as it unfolds throughout an item-writing course, regardless of whether the said development (or lack thereof) happened as a result of the course instruction, of the fact that the participants had several opportunities for item-writing practice they learned from, or of any other factors. Also, it should be kept in mind that the pre- and post-test data are only one segment of the study's dataset, which also includes data from participant interviews and feedback questionnaires. Thus, the pre- and post-test data and their analyses are triangulated, supporting their validity.

3.5.1.2 Pre- and post-test instrument: Item-writing assignment

As explained above, the pretest-posttest design in this study involved an item-writing assignment which was administered to participants before and after they did the training course. Although the item-writing training comprised instruction in writing items to test grammar, vocabulary and the four skills, the assignment was limited to three item types to keep it feasible for the participants: (1) two multiple-choice grammar items targeting the levels A2 and C1 of the Common European Framework of Reference (CEFR); (2) a B2 writing prompt, including a short input text and instructions to produce a formal email in response to the input text; and (3) a B1 listening task, including an input text and a gap-fill task with six items. This was chosen to cover as wide a range of items as possible – in terms of target proficiency levels (A2-C1), target constructs (grammar, one receptive and one productive skill), and task types (selected response, short and long constructed response) – within a manageable timeframe.

The assignment instructions (see *Appendix 4*) included a general introduction that explained the assignment's aim, completion timeframe, and submission instructions. This was followed with detailed guidelines for each item type: item specifications, a sample item, and an item template. Two additional documents were also provided: The *Core Inventory for General English* to be used for selecting grammar exponents, functions, and topics; and a tutorial on using *Lextutor* to check items for vocabulary frequency. Prior to use in this study, the assignment instructions were moderated for clarity by Lancaster University's Language Testing Research Group. The assignment was then piloted with Cohort 2, resulting in minimal revisions (slightly altered layout and changed font colour of several sentences in the instructions).

The participants were instructed to complete the assignment within one week at a place and time convenient to them. They could choose to write all items in one go or over several sittings. In either case, the participants were instructed to record the time it took them to complete each task. The assignments were submitted to the researcher via email immediately upon completion. The items were then labelled, collated, and stored in a secure computer. Before the items were distributed to expert judges to evaluate the item quality, they were anonymized and randomised, with a unique number assigned to each item.

3.5.1.3 Participants: Item-writer trainees

Similar to the pilot study, all participants in the main study were British Council China employees. The participants were self-selected from among those working in the organisation’s four branches in China. Participants worked primarily as language examiners and, as a pre-condition for such role, held a university degree, an English teaching qualification, and a minimum of two years relevant teaching experience. Their contract stipulated that they could be drawn upon for various assessment-related projects of the organisation in the East Asia region, thus the need for the participants to upgrade their language assessment skills, including in item writing. Thirty-five employees signed up for Cohort 3 of the item-writing course; 25 of these completed all course requirements so their data was included in the main study. Out of the ten participants who did not complete the course, three never started it for various personal reasons, while the other three dropped out during the first module of the course having found the workload excessive. The remaining four dropped out mid-course: two because they had resigned from the organisation, and the other two explained their decision with increased work pressure because of weekly trips.

All 25 participants completed a detailed background questionnaire (*Appendix 5*) consisting of seven sections and 21 questions which asked about participants’:

1. *Biodata:* name, gender and age.
2. *Languages:* participant’s L1 and proficiency in any other languages they know.
3. *Educational background:* university degrees and language teaching qualifications.
4. *ESL/EFL teaching experience:* teaching experience, including the number of years and courses taught.
5. *Writing experience:* whether participants had produced written work for publication or to be read/used by others.
6. *ESL/EFL testing experience:* experience of classroom and large-scale assessment.
7. *ESL/EFL test writing experience:* whether participants had produced test items for classroom use or large-scale assessment, and whether they had received any item-writing training.

Six of the 25 participants were female and 19 male. Their age ranged from 29 to 60 years old (M=40.4). They were English-L1 speakers from Australia, Canada, Ireland, New Zealand, the

UK, and the USA, apart from two highly proficient L2-speakers of English (Dutch-L1, Polish-L1). All but one participant spoke minimum one (maximum four) languages other than their L1. Chinese (Mandarin) was the most popular language (20); other languages included Spanish (9), German (5), French (4), and Japanese (4) – 14 different languages altogether. All participants held a Bachelor's degree, with 16 participants also a Master's degree from a variety of subjects, and one a PhD. All participants had a CELTA qualification or equivalent, and four had a DELTA. Their EFL teaching experience ranged between 3-17 years (M=8.5); they had taught in a minimum of one (maximum five) different countries in Europe, Asia, and/or Americas. All had experience teaching general English to adults, many also taught English to young learners (22), exam preparation classes (21), and business English (16). Nineteen participants had no experience writing for publication, while six had published magazine articles, poetry, or ESL/EFL teaching materials. Sixteen had written unpublished materials mostly related to teaching English (e.g. teacher training, teaching and exam preparation). Everyone had some classroom assessment experience as well as having worked as a speaking and/or writing examiner for a large-scale exam board. No-one had produced items for a professional exam board, but ten participants reported that they used to create tests for their former school, university or language centre.

3.5.1.4 Expert judgements

Language testing heavily relies on expert judgement in all aspects of the field: judgements are customarily used to determine test method and content, to create item specifications and scoring rubrics, to rate writing and speaking responses, and to determine item difficulty and cut-off scores (Alderson, 1993). The use of expert judgement as a research method has been described in studies investigating test content (Alderson & Kremmel, 2013; Alderson & Lukmani, 1989; Bachman et al., 1996; Weiler, 2018) and item difficulty (Bachman, 2002; Fulcher, 1997; Hamp-Lyons & Mathias, 1994; Sydorenko, 2011), among others. Although the method is regularly applied, there have been extensive debates regarding its usefulness because judges' agreement levels are often quite low. For example, Bejar (1983), in striving to lower item production costs and limit exposure to items, investigated the possibility of using expert judgements, instead of item piloting, to determine test item difficulty. The study results were disappointing in that, even after training, judges' agreement was unacceptably low. Therefore, Alderson (1993) suggested that "empirical data [should] … replace judgements" (p.47) whenever possible.

Several suggestions have been made on how to overcome this serious limitation of the method. Among them, training judges and using discussions to boost agreement have been proposed. For example, despite Bejar's (1983) evidence to the contrary, some studies (e.g. Fortus et al., 1998; Hambleton & Jirka, 2006; Lumley, 1993) found that judges' agreement improved substantially after a training session. Similarly, judges in Fulcher's (1997) study first worked independently and then gathered for a face-to-face session where they were given an opportunity to discuss their judgements and amend the ratings they had provided. Inter-rater reliability was calculated for both sets of judgements and the ones obtained after the discussion proved substantially higher. However, Alderson (2010) and Alderson and Kremmel (2013) argue against training judges because such training "would amount to cloning" and, instead of drawing on judges' expertise, the study results "would simply indicate the success of the cloning process" (Alderson, 2010, p.96). Using discussions as a means of reaching agreement is also deemed unacceptable because "a forced agreement through discussion" (Alderson & Kremmel, 2013, p. 538) would only be a continuation of the cloning process. The caution against reaching agreement through discussion has been supported with research into rater judgements: van Moere (2014) described that, when rating as a group, rater personalities and seniority might influence the rating outcomes. Also, raters might engage in a "trade-off" (p.1362), that is, if one rater's opinion prevailed on the last candidate's performance, s/he will let others win on the next one. All the above-mentioned considerations influenced my decision to not train the expert judges in this study, as well as to let the judges work independently instead of collaboratively.

Despite the expert judgement method being seen as problematic, Alderson (1993) concedes that sometimes expert judgement can be unavoidable "in the absence of other data" (p.56). One of the areas where expert judgement has, until present, found no substitute, is item quality review. To the best of my knowledge, all large-scale exam boards use expert reviewers to control the quality of test items as part of the item production process. To illustrate, I will provide very brief accounts of the item review procedures for the IELTS test by Cambridge Assessment, the TOEFL iBT test by ETS, and the Aptis test by the British Council.

The IELTS test production process is outlined in the brochure *The IELTS Question Paper Production Process* which, until recently, was available on the IELTS website (www.ielts.org). It describes two stages of item review: *pre-editing*, during which first drafts of submitted items are scrutinised, and *editing*, when items are reviewed again after revision. Expert judgement is the cornerstone for both stages, the experts being "chairs and Cambridge ESOL staff" (IELTS, 2007, p.2) who review the submitted material during face-to-face meetings. Judgements are made on the "topic, topicality, level of language, suitability for the task, length, focus of text, style of writing, focus of task, level of task" (IELTS, 2007, p.2). In other words, expert judgement is relied upon in all aspects of test item quality review. It should be noted that discussion is used as a way of disagreement resolution, and item writers are encouraged to attend the meetings which are seen as an opportunity for item-writer professional development (Green & Hawkey, 2012).

The TOEFL iBT test production cycle contains three types of item review: content review, fairness review, and editorial review. At the content and fairness stages, judgements are made by experts who are "assessment specialists" (TOEFL iBT, 2018, p.8), while editorial review is done by professional editors. As admitted by ETS themselves, "[e]xpert judgement… plays a major role in decision whether a Speaking or Writing item is acceptable and can be included in an operational test" (p.8). Unlike for IELTS, item review for TOEFL iBT happens sequentially with four or more reviewers taking part in the process. They can consult each other, though, and items are accepted only "if all reviewers judge them to be acceptable" (p.7).

The Aptis production cycle similarly includes a quality review stage whereby "[i]tems are annotated by two independent reviewers, using a number code system. This identifies any element of the item that does not meet any part of the specifications" (Aptis, 2015, p.30). Two details are of interest in the cycle's description: the review is done independently by each reviewer and, by introducing the "number code system", an attempt is made to 'automatize' or 'objectivize' the review process thus making it less dependent on subjective human judgement. However, in essence, human judgement remains at the heart of the review process.

It should be noted that item production cycles for large-scale examinations also include an item piloting stage where items are tried out on a representative test-taker sample to predict how they would function in a live test. For example, at Cambridge Assessment and as reported by Saville (2003), writing and speaking prompts undergo trialling "with small but representative groups of candidates" (p.90), while item-based tasks are pre-tested on large groups of candidates "which then allows the items to be analysed statistically" (p.90). One important aim of piloting is to determine item difficulty, but it can also uncover item deficiencies that went unnoticed during the editing process (Green & Hawkey, 2012). Although piloting is an important way of determining item quality, it was decided not to employ it in the present study. From a practical perspective, piloting would be unfeasible because of the large volume of items involved: test-takers would have to complete 50 listening tasks, respond to 50 writing prompts, and do 100 multiple-choice grammar items. It would also be unclear how to choose a representative test-taker sample as the items were not produced for a specific test. Moreover, such an immense undertaking, if at all possible, would not result in information sufficiently valuable for the present research project to justify the effort because the aim of this study is to explore item-writing skill development resulting from training and not to identify, with precision, those items that would be unfit for inclusion in a live test.

Participants: Expert judges

It seems obvious that, for the expert judgement method to provide valid (if not always reliable) results, suitable individuals are to be employed as expert judges. It is then somewhat surprising that, although expert judgement is widely used in language testing, no clear position exists on who is to be considered an expert in the field. The studies mentioned in the previous section either made no reference to how judges were identified or provided a one-sentence rationale for judges' selection. For instance, Alderson and Kremmel's (2013) participants were "involved in language test development at a national or institutional level" as well as "teaching at a university level" (p.541), while Bachman et al. (1996) employed "trained applied linguists with experience as EFL teachers and/or administrators" (p.131). It seems that the authors of the former study considered formal education in the relevant field an indication of expertise, while in the latter study professional experience was valued more. Among the nine empirical studies discussed in the Literature Review Chapter (see Section 2.2.7.1) that investigated item-writing training effectiveness using the expert judgement method, only two provided some (limited) information about the judges. Dellinges and Curtis (2017) stated that the judges were two faculty members, one with a doctoral degree in education and the other a trained psychologist. Scott et al., (2019) mentioned that the two judges were "educators" and "item-writing experts" (p.12); however, no justification was provided as to why the judges were considered experts.

At the same time, social sciences have long been concerned with the notion of expertise, and a number of studies have looked into the nature of expert knowledge. The seminal volume *The nature of expertise* edited by Chi, Glaser and Farr (1988) provides eight characteristics of expert knowledge (Glaser & Chi, 1988). Among those, one has proved particularly instructive for the purpose of the present study, namely that experts excel only in their narrow domain and "there is little evidence that a person highly skilled in one domain can transfer the skill to another" (Glaser & Chi, 1988, p.xvii). The implication for this study is that expert judges were not to be selected on the basis of having a degree in applied linguistics or even language testing; or because they are experienced language teachers; or because they have experience working as raters or even item writers – the judges had to be item reviewers with years of professional practice and training in reviewing items against detailed specifications.

Another important decision concerned the number of judges in the study. The overview of item-reviewing practices at large-scale examination bodies revealed that any number of reviewers, from two (Aptis) to four (TOEFL iBT) and possibly more (Cambridge Assessment) can be involved in the editing process. The research literature is similarly conflicting. For instance, Spaan (2007) writes that items should be reviewed by "an individual or a small group of experienced item writers" (p.282), Osterlindt (1998) states that it is not the number of judges but their level of agreement that is important (p.260), while Bejar (1983) concluded that robust agreement is only possible when no fewer than 20 judges are employed, which is unrealistic even for an organisation like ETS. The four empirical studies discussed in the Literature Review Chapter (see Section 2.2.7.1) that investigated item-writing training effectiveness and reported on the number of judges, used either one (Naheem et al., 2012) or two judges (Dellinges & Curtis, 2017; Scott et al., 2019; Yurdakul et al., 2020).

For the present study, two important issues had to be considered when deciding on the number of judges: (1) individuals with the required level of expertise are few, sought after, and normally well paid; and (2) there was a large number of items for review, amounting to ten days of full-time work. To provide judges with reasonable renumeration, I applied for the competitive British Council Assessment Research Award scheme for doctoral studies support, with a successful outcome. The award allowed me to employ two judges, who were identified among experienced item reviewers (with at least three years of item reviewing experience working for a large-scale examination body) known to me and/or my supervisor. Potential judges were approached via email with an invitation to take part in the study. They were provided with an information sheet that described the aims of the study, potential benefits, disadvantages, as well as the ways the judges' identity is protected. The two reviewers who agreed to take part in the study signed a consent form.

However, I felt an odd number of judges was needed. This is because there was a potential for the two judges to award diverging scores (e.g. '1' and '2') but, to preserve the original three-band evaluation scale, decimal points in item evaluations had to be avoided (see a more detailed discussion later in this section). That is why I, the researcher, acted as the third judge in this study. My own professional item-reviewing experience made me suitable for this role. It should be noted that my role was as a third independent judge and not as an arbiter of diverging scores for judges 1 and 2. However, I also acted as one of the tutors for the item-writing course and there was a possibility that my judgements could be affected if I remembered some of the items I had seen before. Therefore, after the items were anonymised and randomised, I made sure at least half a year had passed before I started the evaluation process, for my memory of the items to fade.

Expert judgement instruments and procedures

An evaluation scale (see *Appendix 6*) was developed for the expert judges to assess the quality of items produced by the item-writer trainees. Evaluation criteria for the first version of the scale, used during the pilot study, were derived from the item specification requirements. The specifications themselves, as explained in Section 1.2.2, were informed by the test development principles formulated in the socio-cognitive framework (O'Sullivan & Weir, 2011; Weir, 2005). For example, the socio-cognitive framework follows the cognitive processing approach to listening sub-skills' acquisition (Field, 2019) whereby the acquisition is seen as a progression from lower-level processing (decoding of individual sounds and words) through parsing (producing abstract propositions of messages using one's own words) to higher-level processing (inferencing, applying world knowledge, and constructing discourse). The listening task specifications for the pre/post item-writing assignment (*Appendix 4*) were created with B1 proficiency level in mind, therefore items were to target mid-level processing (parsing). To achieve this, item stems were to paraphrase the information heard in the text. Subsequently, Evaluation Criterion 10 of the listening task evaluation scale reflected this specification requirement: "Stem: a paraphrase, i.e. does not literally repeat what is heard in the text" (*Appendix 6*).

Another aspect afforded much attention in the socio-cognitive framework is situational and textual authenticity of test tasks. In terms of situational authenticity, to continue with the listening task example, each listening input text was to be situated within a plausible context, with the speaker, the situation, and the purpose of listening specified in the instructions to test-takers (Weir, 2005). This requirement was included in the listening task specifications (*Appendix 4*), and then reflected in the listening task evaluation scale under Evaluation Criterion L18: "Instructions include all specified information – the speaker, the situation, guidance in how to fill the gaps" (*Appendix 6*). In terms of textual authenticity, the listening task specifications stated that the input text had to "sound like authentic spoken English and not a written script read out" (*Appendix 4*). In line with this specification requirement, the evaluation scale for the listening task included Evaluation Criterion L14: "text sounds authentic according to the genre" (*Appendix 6*).

The evaluation scale included 15 evaluation criteria for grammar items, 12 for writing prompts, and 21 for listening tasks. Each criterion was scored on a three-band scale: '2' - the item fully conforms to the specifications on this criterion; ‘1’ - some improvement is needed; ‘0’ - the item failed to conform to the specifications on this criterion. Detailed band descriptors were developed for each band of each criterion and, where appropriate, questions were included to guide the reviewers during their work. To exemplify, Evaluation Criterion 2 for three-option multiple-choice A2/C1 grammar items focussed on the quality of distractors. The band descriptors for the criterion were as follows:

- ***Band 2:*** *both distractors are strong*
- ***Band 1:*** *one of the distractors is weak*
- ***Band 0:*** *both distractors are weak*

To further guide the reviewers in their decisions, supporting questions were provided. The following questions illustrate these for Evaluation Criterion 2:

- *Are the distractors plausible?*
- *Will students who have mastered the grammar point tested have more chance to answer correctly?*
- *Will the distractors work well in differentiating between weak and strong students?*
- *Is it possible to discard any of the distractors without having mastered the grammar point tested?*

The evaluation scale was trialled during the pilot study with the course tutor as a judge. The trial revealed that some flaws in trainees’ items were not picked-up by any of the evaluation criteria. This situation is not unknown in operational item writing and, from my personal experience working as an item reviewer, Quality Review (QR) sheets are regularly updated based on item-reviewer feedback, to include item problems unforeseen in the specifications. Therefore, two types of changes were made to the evaluation scale:

1. Additional evaluation criteria were included, based on the pilot study item review;
2. An Overall Item Acceptability Criterion was added for each item type, asking the reviewer to judge the overall acceptability of the item for inclusion in a live test. The reason for this was twofold: it provided the judges with an opportunity to look at an item as a whole instead of only scrutinising its minute characteristics, and it also served as a safeguard in situations when an item flaw was not included in the evaluation criteria.

After these revisions, the final version of the evaluation scale contained 19 criteria for grammar items, 17 for writing prompts, and 27 for listening tasks. A substantial number of criteria were 'mechanical' in nature, for instance the number of words in a text / an item stem, the frequency of item lexis, and so on, and could be 'objectively' assessed (e.g., by counting words or running a word frequency check with *Lextutor*). Therefore, all objectively-scored criteria were removed from the evaluation scale intended for expert judgement, thus reducing the experts' workload and allowing them to concentrate on subjectively-scored criteria that require human judgement (e.g. judging authenticity of an input listening text or appropriacy of its content) . Instead, scores on objectively-scored criteria were calculated by the researcher twice to exclude the possibility of human error, and any discrepancies resolved.

A total of 280 items were collected across the pilot and the main study (Cohorts 2 & 3), as shown in *Table 3-1,* and the judges were requested to provide evaluations of all 280 items using the final item evaluation scale, although the data from the pilot study was excluded from the main study analyses. The reason for the inclusion of pilot study items in the item evaluation process concerns the assumptions for the Intraclass Correlation Coefficient (ICC) test, which was used to determine agreement between judges - see Section 3.5.1.4.

Table 3-1. Items produced for the item-writing assignments

	A2/C1 grammar items	**B2 writing prompts**	**B1 listening tasks**	**Total per study**
Pilot study (pre & post)	40	20	20	80
Main study (pre & post)	100	50	50	200
Total per item type	140	70	70	280

One week prior to the commencement of the evaluation work, the judges were sent the evaluation scale, an Excel spreadsheet to enter the scores, and item review guidelines. The judges were encouraged to study the documents and, if necessary, ask for clarification. A week later, they were sent items for review. The items had been anonymised and randomised so that the judges would not know which trainee each item was written by, whether it was written before or after the training, or whether it came from Cohort 2 (pilot study) or Cohort 3 (main study) of the course. The judges could work at a place and to a schedule convenient for them but had to submit the evaluations via email within one month of receiving the items.

Expert judgement analyses

The judgement data analysis began after all experts had submitted spreadsheets with item evaluations. To prepare the data for analysis, the evaluations were de-randomised and de-anonymised to match each item score to: a) the participant who produced the item, b) the pre- or post-training assignment, and c) the cohort the item came from. Four types of quantitative analyses were then conducted: (1) judges' agreement was established through Intraclass Correlation Coefficients; (2) descriptive statistics of item evaluations were calculated and interpreted separately for each item type; (3) Wilcoxon signed-rank tests were performed on raw item evaluation scores to examine the statistical significance of changes in item quality from before to after training; and (4) gain ratios were calculated to explore individual item-writer variation.

1) Intraclass Correlation Coefficient

Intraclass Correlation Coefficients (ICC) were calculated to determine agreement between the judges. The use of ICC in assessing consistency of measurements by different raters has been widely described in the literature (e.g., McGraw & Wang, 1996; Shrout & Fleiss, 1979). ICC is particularly recommended for measuring homogeneity of ratings when the analysis concerns not only pairs of ratings but "larger sets of measurements" (McGraw & Wang, 1996, p.30). Cicchetti (1994) argues for the usefulness of ICC over other reliability measures because

> …it distinguishes those sets of scores that are merely ranked in the same order from test to retest from those that are not only ranked in the same order but are in low, moderate, or complete agreement with each other; and … it corrects for … agreement expected on the basis of chance alone (p.286).

To use ICC appropriately, the correct form had to be selected from the eight forms that exist, with three factors to take into consideration: model (one-way random effects, two-way random effects, two-way mixed effects), type (single or average measures), and definition (absolute agreement or consistency). After considering the type of data and the study's design, the two-way mixed-effect average measures model based on absolute agreement was selected, because each item was evaluated by each of the three judges, who were the only judges in the study (Shrout & Fleiss, 1979, p.421). *Average measures* were determined as it is judges' agreement on average ratings per criterion that is of interest, and *absolute agreement* was identified because the analysis was concerned with whether different judges assigned exactly the same score on the same criterion, and not merely whether judges' ratings were consistent with each other.

Another important consideration concerned the size of the dataset. A low ICC can stem not only from low agreement among judges, but also from a small number of subjects and/or judges in the study. Koo and Li (2016) recommend, as a rule of thumb, “to obtain at least 30 heterogeneous samples and involve at least 3 raters” (p.158) for ICC results to be meaningful. Taken separately, the two cohorts of this study do not conform to this requirement, but put together they have 35 participants, so the requirement is met. This is why main study judgements on items from the pilot study were included in the expert judgement strand of the Pretest-Posttest study.

ICC was calculated using SPSS 25 for each criterion, separately for the items produced pre- versus post-training. The output was interpreted using guidelines by Cicchetti (1994), which are widely cited in the literature: ICC below .40 means ‘poor agreement’, between .40 and .59 – ‘fair’; between .60 and .74 – ‘good’, between .75 and 1.00 – ‘excellent’ (p.286). Instances of poor, fair, good and excellent agreement were counted and compiled in *Table 3-2*.

Table 3-2. Judges’ agreement: ICC test results

		Poor agreement	**Fair agreement**	**Good agreement**	**Excellent agreement**
Grammar A2 items	Pre-training	4	2	3	0
	Post- training	6	1	0	1
Grammar C1 items	Pre- training	2	2	4	1
	Post- training	5	3	1	0
Writing B2 prompts	Pre- training	2	6	3	1
	Post- training	4	5	2	1
Listening B1 task	Pre- training	5	7	1	2
	Post- training	11	2	1	1

Overall, the agreement among judges, as shown in *Table 3-2,* was found to be quite low. In the table, the figures represent instances of each type of agreement per item type, based on agreement for individual criteria. For example, for grammar A2 items produced for the pre-training assignment, judges demonstrated poor agreement on four criteria and fair agreement on two criteria. It follows from the table that instances of poor and fair agreement were quite numerous, while excellent agreement was rare. Interestingly, the agreement on post-training items for all item types diminished compared to pre-training, with more instances of poor agreement and fewer instances of good agreement.

This generally poor agreement among judges is, however, not surprising given multiple previous studies reporting similar results (see e.g., Alderson & Lukmani, 1989; Bejar, 1983; O'Neill et al., 2019). There can be several explanations for the low agreement. Firstly, as discussed above, low ICC can be the result of a small sample size because, with three judges and 35 participants, this study is a borderline case. Secondly, the judges received no training and could not discuss their ratings with each other to increase the chances of agreement, as explained earlier in this section. This generally low reviewer agreement that tends to occur whenever reviewers work independently might be the reason why some exam boards have editorial meetings (e.g. Cambridge Assessment) or encourage consultations among reviewers (e.g. ETS). However, because the aim of the item evaluation in this study was not to standardise ratings but to draw on judges' extensive expertise as item reviewers, low agreement is not necessarily a weakness of measurement in this study. Although the judges were using the same evaluation scale, each judge may have looked at the items from a slightly different perspective and may have noticed slightly different things.

Analysing three separate item evaluation datasets was, however, methodologically unfeasible –three independent sets of analyses would have to be conducted each of which would then have to be integrated with findings from the qualitative interview data. It is also unclear how findings from the analyses of three independent datasets could then be integrated for the final analysis. Therefore, for the purposes of further analyses and with the view of obtaining a dataset that would best reflect each judges' opinion, it was decided, for all subjectively-scored criteria (the criteria that were evaluated by the three judges) to use the average of the three judgements to create a so-called 'final item evaluation dataset' which would then be used to explore the effect of training on item quality. The median score was used to establish the final evaluation for each criterion of each individual item, based on the principle of commonality. For example, if two judges assigned band '1' and one judge assigned band '2' on a specific criterion, '1' was used in the Final Dataset. The median score was also preferred over the mean score to avoid decimal points in item evaluations, thus preserving the original three-band evaluation scale.

The scores on the objectively-scored criteria (the ones that were calculated by the researcher) were added to the Final Dataset. Lastly, three types of total scores were calculated for each item and added to the Final Dataset: (1) the sum of scores for all criteria on which the item was evaluated; (2) the sum of scores for all objectively-scored criteria on which the item was evaluated; (3) the sum of scores for all subjectively-scored criteria on which the item was evaluated. The decision to calculate total scores separately for objectively-scored and subjectively-scored criteria was made because of the substantial differences between the two types of criterion and also because a different method was used to arrive at the scores for each type of criterion. It must be mentioned that the Final Dataset comprised the scores for Cohort 3 (main study) items only. This is because Cohort 2 and Cohort 3 are not fully comparable as the course evolved with each run, including changes to the course structure, training materials and activities (see Section 3.2). The Final Dataset was then used to perform a range of statistical analyses, as described below.

2) Descriptive statistics

Descriptive statistics were calculated for: a) the sum of scores for all criteria of each item; b) the sum of scores for the objectively- and subjectively-scored criteria of each item; c) the scores for each individual criterion of each item. More specifically, the following descriptive statistics were calculated using SPSS 25: range, minimum, maximum, mean, standard deviation, skewness, and kurtosis (including the standard error where appropriate). The statistics were interpreted separately for each item type.

3) Wilcoxon signed-rank tests

Wilcoxon signed-rank tests were used to examine the significance of changes in raw item evaluation scores from before to after the training. The tests were run for each individual criterion, as well as the sums of scores. This non-parametric test was selected because the item evaluation data for each individual criterion is ordinal. However, score sums form an interval scale. Therefore, to determine which statistical test to use – parametric or non-parametric – the assumption of normality was tested (McCrum-Gardner, 2008). Two tests of normality available in SPSS25 were performed: Kolmogorov-Smirnov and Shapiro-Wilk. In cases when contradictory results were obtained, the Shapiro-Wilk test was preferred as having better power compared to the majority of tests of normality (Yap & Sim, 2011). The tests of normality demonstrated that the majority of score sums did not meet the assumption of normal distribution (*Appendix 7*). In five instances, the assumption of normality was met; however, the other member of the pair did not meet the assumption, therefore Wilcoxon signed-rank tests were still the right choice. In one instance – the sum of scores for B1 listening tasks – both pre-and post-training data were normally distributed allowing for a paired-samples t-test to be performed. However, to maintain consistency of the statistical measure across item types, and because a non-parametric test produces more conservative results, the decision was made to perform Wilcoxon-signed rank tests on all item evaluation data.

Z-scores, asymptotic significance values (*p*), and effect sizes (*r*) were calculated. SPSS does not calculate effect size, so it was obtained manually based on the equation recommended by Field (2013, p.234) which was preferred as more rigorous compared to other formulae in the literature (see e.g., Rosenthal, 1991). The interpretation of results was as follows: if the values were statistically significant based on positive ranks (i.e. item scores were significantly higher post-training compared to pre-training), the quality of items produced post-training (with respect to the criterion under consideration) was significantly higher than the quality of items, on the same criterion, produced pre-training. The significance level (*p* value) was set at below .05, which is typical for social sciences (Wheelan, 2013). The effect size was interpreted using Cohen's (1988) guidelines: values ≤ 0.3 are viewed as a small effect size, 0.3 to 0.5 represent a medium effect size, and ≥0.5 a large effect size.

4) Gain ratios

Although Wilcoxon signed-rank tests produced important comparative statistics on the judged quality of items from before to after the training, the technique has its limitations. Firstly, the item quality for each criterion is averaged for all course participants, therefore, no individual differences among participants can be detected, while this study is interested not only in the cohort of trainees as a whole, but also in the trajectory of item-writing skills development of each individual participant. Secondly, it can be misleading to compare pre- and post-test results using Wilcoxon signed-rank tests in situations where pre-test scores are already quite high. The latter was the case for some evaluation criteria in this study, whereby items received high scores on the criterion before the training. This made the change in quality post-training, though meaningful, not numerically large, so the change passed undetected by Wilcoxon signed-rank tests. Therefore, the pre- and post-training item scores were also examined using a gain ratio technique - a "more informative value through which to view the change" (George & Cowan, 1999, p.69).

Analysing gain ratios is viewed as the preferred technique when evaluating the influence of instruction on learning (George & Cowan, 1999). For example, the maximum total score for a grammar item in this study, as the sum of scores on 18 individual criteria, is 36 (excluding the Overall Acceptability Criterion). Let us consider a case where participant A's grammar item received only 10 points pre-training, but the participant wrote a much better grammar item post-training which scored 25. On the other hand, participant B's pre-training grammar item already scored very high, 32 points, and his post-training item reached the maximum score of 36 points. The absolute gain for participant A is 15 points while participant B's absolute gain is

4. The question arises whose improvement is more meaningful. Participant A's item is much better post-training; however, it is still not good enough to be used in a live test because it did not achieve band '2' on all criteria. At the same time, participant B's item, in terms of absolute gain, is only marginally better post-training, but this small difference is of vital importance because it made the item acceptable for operational testing. If we calculate gain ratios, the statistic will reflect this important change much more accurately. Participant A's post-training item gained 15 points out of 26 possible[4]: 15 ÷ 26 = .58, that is the gain ratio is 58%. Participant B's post-training item gained 4 points out of 4 possible: 4 ÷ 4 = 1.0, that is the gain ratio is 100%. In terms of gain ratios, as well as in real practice, the second result is distinctly better than the first.

To investigate changes in the quality of items from before to after the course for each individual participant, gain ratios were calculated for the sum of scores on all criteria for each individual item, as well as the sums of scores on the objectively- and subjectively-scored criteria. The gain ratios were calculated manually using the technique described above. The results were analysed separately for each item type, to provide insights into more nuanced changes in the quality of items from before to after the training, as well as into item-writer variation within the training cohort.

3.5.1.5 Interviews

To obtain qualitative data about participants' experiences of writing items for the two item-writing assignments, participants were interviewed. As noted in the Literature Review Chapter (Section 2.2.7.2), compared to a large volume of research into rating processes, research into item-writing processes is scant. To the best of my knowledge, only three studies are available to date in the field of language testing. Kim et al. (2010) involved four item-writer participants who were interviewed as well as kept diaries about their item-writing experiences. Green and Hawkey (2012) engaged seven item writers with whom they conducted individual interviews, recoded a focus group discussion, and observed an item-editing meeting. Salisbury (2005) conducted the most comprehensive of the existing studies, with ten item writers, using a combination of think-aloud and interview methods. There also exist two item-writing studies from fields other than language testing: Fulkerson and Nichols (2010) investigated item-writer cognitive processing by combining think-aloud and interview methods. Three item writers were provided with ready-made scenarios to create two MCQs for a science test. Johnson et

[4] The possible gain is the difference between the maximum score a post-training item could achieve and the score the pre-training item achieved. In our example, it is: 36 (the maximum score) – 10 (the score for participant A's pre-training item) = 26.

al. (2017) looked into the item-writing processes of seven item writers of GCSE tests from different subject areas (Biology, Geography, Mathematics, and Physics). The study employed video observation and interview methods. *Table 3-3* summarises the data collection methods used in the studies.

Table 3-3. A summary of research methods used to investigate item-writing

Data collection methods	Kim et al. (2010)	Green and Hawkey (2012)	Salisbury (2005)	Nichols and Fulkerson (2010)	Johnson et al. (2017)
# participants	4	7	10	3	7
Diary	*				
Focus group		*			
Interview	*	*	*	*	*
(Video) observation		*			*
TAP[5]			*	*	

These data collection methods were considered while choosing the most suitable method for the present research project. Four of the five methods presented in *Table 3-3*, namely diaries, observations, focus group interviews, and TAPs were deemed unsuitable, for the following reasons:

- In the study by Kim et al. (2010), item writers kept diaries over an extended period of time, while the pretest-posttest design adopted for this study presupposes data collection at two discrete points in time, thus excluding the diary method.
- The present study's aims required data to be collected individually from each participant to analyse their individual item-writing processes and strategies, which necessarily excludes the focus group method.
- The observation method could not be implemented because (1) the item-writing event required an extended period of time and often happened over several sittings, and (2) the 25 participants in the study did the assignments at a time and place that suited each of them individually.

[5] Think-aloud protocols

- Concurrent TAPs are intended to yield factual reports about the content of the process (Taylor & Dionne, 2000), whereas the present study is also interested in the participants' perceptions about item writing they did.

Moreover, it has been noted in the literature that TAP might change the nature of the process under investigation (Barkaoui, 2011; Leighton, 2017). Creating test items requires full concentration and constant attention to the specifications, while the requirement to provide an ongoing commentary would necessarily disrupt the process, especially for novice item writers who have not yet established item-writing routines. Another, practical consideration should also be noted. The two item-writing studies that employed TAP methodology (Nichols & Fulkerson, 2010; Salisbury, 2005) were small-scale studies involving a maximum of ten participants and one item type. The assignment in this study included four items, three item types, and 25 participants, 17 of whom were interviewed pre-training and 19 – post-training. Thus, the sheer scale of the study renders TAP methodology unfeasible.

Interviewing has so far been the most popular data collection method in this topic area as it was used in all five studies (*Table 3-3*), while each of the other four methods was used in one or two studies only. This preference for interviews might be explained by the nature of the item-writing process. In real-life settings, an item is rarely written in one sitting, with the item writer returning to it several times for revision until s/he is satisfied that the item conforms to all specification requirements. However, even if an item-writing event happens in one sitting, it can take longer than a TAP or an observation can afford. Another strength of the interview method is that it allows the researcher to elicit more information from unforthcoming participants through probing, asking for clarification, or offering follow-up questions (Johnson, 2002). The retrospective interview method does, however, come with its own drawbacks, the biggest of which is the risk of recall bias and researcher bias, which might distort the data (Green, 1998). However, having considered all possibilities, it was concluded that the retrospective interview method was the most suitable for this research study.

Of the three major interview types – structured, unstructured, and semi-structured – the latter was deemed as most suitable for the study, since "[s]emi-structured interviews are used when the researcher knows enough about the topic or phenomenon to identify the domain … but does not know and cannot anticipate all of the answers" (Morse, 2012). Indeed, although the item-writing domain was already familiar to me as a practicing item writer as well as a language testing researcher, it was impossible to fully predict participants' understanding of the item-writing process, the approaches they might have taken, or the strategies they might have developed. Therefore, it was important to let interviewees shape the interviews, in particular to allow for "unexpected themes to emerge" (Mason, 2004, p.1020). However, because of the retrospective nature of the interviews that asked the participants to recollect a recent experience, an 'aide memoire' (Mason, 2004) was needed: the item-writing assignment including the items the interviewee had produced acted as such. The next section describes the materials and procedures that were used to conduct the interviews.

Interview instruments and procedures

An interview protocol (*Appendix 8*) was developed to conduct retrospective interviews with participants after completion of their item-writing assignment. Based on recommendations for best interview practice (Creswell, 2013; Kvale & Brinkmann, 2015), the protocol included three stages: (1) a *'before the interview'* stage where the research aims were introduced, the interview procedures explained, and confidentiality ensured; (2) an '*interview'* stage which contained a question schedule; (3) an *'after the interview'* stage which included thanks and closing remarks. For the interview stage, a semi-structured interview schedule was developed consisting of main and follow-up questions. Main questions were open-ended and broad in nature to allow interviewees as much freedom in their responses as possible. The aim was to fully elicit information the interviewees were willing to supply by not prompting or leading them unnecessarily (Rubin & Rubin, 2005). The final question "*Is there anything else you would like to tell me about your item-writing experience?*" also aimed to ensure that each interviewee would have a chance to talk about anything else related to item-writing that they wanted. The schedule also included follow-up questions for deeper probing by the interviewer if the initial responses were (too) short or superficial.

The questions formulated for the pre/post training interview schedule reflected my initial assumptions about item-writing skill development (see Section 1.2.1). It was assumed that after the training - as compared to before it - participants would produce better quality items in a shorter period of time, participants would report a better-organised item-writing approach characterised by the use of item-writing strategies, and would be better able to articulate this approach. It was hoped that the interview questions would elicit data to answer RQ2 and RQ3 of this study:

RQ2: How did the participants' item-writing skills develop through the training, as perceived by the participants in interviews?

RQ3: What role did the participants perceive the training played in their item-writing skill development?

The first part of the interview schedule asked interviewees about their experience writing items for the relevant assignment: *"Can you tell me about your item-writing task? How did it go?"* The interviewees were encouraged to talk about the time it took them to do the task, the way they organised their work, and the approach they took to write each item. Because identical questions were included in the pre-course and post-course interview schedule, it was expected that a comparison of pre- and post-training responses would demonstrate that participants had spent a shorter time in producing items after the training compared to before it; that some item-writing strategies had been employed by the participants following the training; and that more participants had been able to clearly articulate their item-writing approach (see Section 1.2.1). The pre/post-course interview schedule also asked participants how easy they found producing the items, and what difficulties they experienced. It was expected that the participants would have found item-writing easier following the training compared to before it and would consequently have reported fewer difficulties. The data elicited through these questions would then be used to answer Research Question 2 of this study.

In the pre-course interview, participants were also asked to speculate on the knowledge and skills they felt they were lacking while doing the assignment. This question reflects Fulcher's (2012) view of LAL as encompassing both theoretical knowledge and practical ability. If it was found that the knowledge/skills reported as lacking by participants were the ones that had been included in the course syllabus, which suggests that the course had met participants' training needs and, therefore, had played a positive role in the participants' item-writing skill development (RQ3). The post-training interview, which included questions about the interviewees' item-writing experience (outlined above), also aimed to elicit reflections on the item-writing training: aspects of the training which interviewees felt were particularly helpful when producing items for their post-training assignment; knowledge or skills that were, in their opinion, missing from the course; their feelings of confidence and readiness to start working as item writers. It was believed that, if participants reported many aspects of the course as helpful, this would serve as an indication that the training had played a positive role in the participants' item-writing skill development. If, on the contrary, the participants reported a lack of confidence in item writing and that many of their training needs were not met, this would signify that the course had not been helpful in developing the participants' item-writing ability. This data would help to answer Research Question 3 of this study.

The interview protocol was pre-piloted with a fellow PhD student in the Department of Linguistics and English Language at Lancaster University who did a shorter version of the item-writing assignment. The pre-pilot interview was video-recorded and analysed for (1) whether the interview protocol was functioning as intended, and (2) whether the researcher performed well as an interviewer to create a secure and comfortable environment, build trust, and engage with the interviewee without being overbearing (Rubin & Rubin, 2005). The interviewee's opinions were also sought on how the interview questions and procedures could be improved.

The retrospective interviews were then conducted with pilot study participants: eight Cohort 2 participants were interviewed pre-training and nine post-training. Prior to the interview, participants were instructed to have the assignment instructions and the items they wrote for the assignment to hand. During the interviewing, the items were used as a prompt: the participants were reminded to refer to the relevant item and were allowed time, if necessary, to read through it again. The interviews were conducted via *Wechat* voice call facility, audio-recorded, and transcribed.

No changes were made to the interview protocol following the pilot study as the protocol was felt to be functioning well for its purpose. The main study interview procedures were also identical to the ones used for the pilot study. Out of the 25 participants, 17 volunteered to be interviewed pre-training and 19 post-training. Of those, 16 participants were interviewed both pre- and post-training, one was only interviewed pre-training, and three were interviewed post-training only. Most interviews recorded after the training were somewhat longer than the ones recorded before, the average length being 16'52'' pre-training and 19'51'' post-training. Individual interviews differed substantially in their length both before and after the training. The shortest pre-training interview came from Arthur (10'52'') and the longest one from Daniel (24'23''). After the training, Austin's interview was the shortest at 12'56'', while Mason's interview lasted 29'30''.

Interview analyses

The Grounded Theory approach

The research methodology literature offers a range of analytical approaches to qualitative data analysis. Kvale and Brinkmann (2009), when describing analysis for meaning (as opposed to analysis for language), distinguish between two approaches: content analysis and grounded theory. *Content analysis* is more concerned with data quantification by assigning text fragments to categories and determining “how often specific themes are addressed in a text” (p.203), while the grounded theory approach is qualitative in essence and does not rely on data quantification for analysis. Rapley (2011) provides an account of four approaches to qualitative data analysis: framework analysis, thematic analysis, interpretive phenomenological analysis, and grounded theory. While the first three approaches, according to Rapley’s description, are concerned with generating, refining, organising and explaining themes found in the data, the grounded theory’s ultimate goal is new theory generation. As discussed in the Literature Review Chapter (see Section 2.2.7.2), few studies have focussed on item-writing processes and, to the best of my knowledge, no studies exist that have explored item-writing skill development processes in the way this study does. Given the current absence of item-writing skill development theories, the aim of this study is to generate such a theory through the process of data exploration. Therefore, the grounded theory approach was deemed most suitable for this study.

Grounded theory was first proposed by Glaser and Strauss (1967) as a way of increasing explanatory power in qualitative research. The approach allows researchers to move beyond description by identifying patterns in the data and developing new concepts (Charmaz & Bryant, 2011, p.348). Coding lies at the heart of the grounded theory approach as codes are relied upon to form the basis of an emerging theory. Corbin and Strauss (2015) distinguish three types of coding (see *Figure 3-2*). At the initial *open coding* stage, defined as “breaking data apart and delineating concepts to stand for interpreting meaning of raw data” (Corbin & Strauss, 2015, p.239) the researcher is engaged in re-reading, breaking down, examining and conceptualizing the collected data. Next follows the *axial coding* stage where codes created during the open coding stage are analysed comparatively in order to discover connections between them, to identify categories, and to link “properties and dimensions to codes” (Corbin & Strauss, 2015, p.241). At the final *selective coding* stage categories are integrated to form a theory. Open and axial coding can be done iteratively because they inform each other: during the open coding, the reasoning moves from data to codes while in the axial coding the reasoning moves from codes to data (Boeije, 2010).

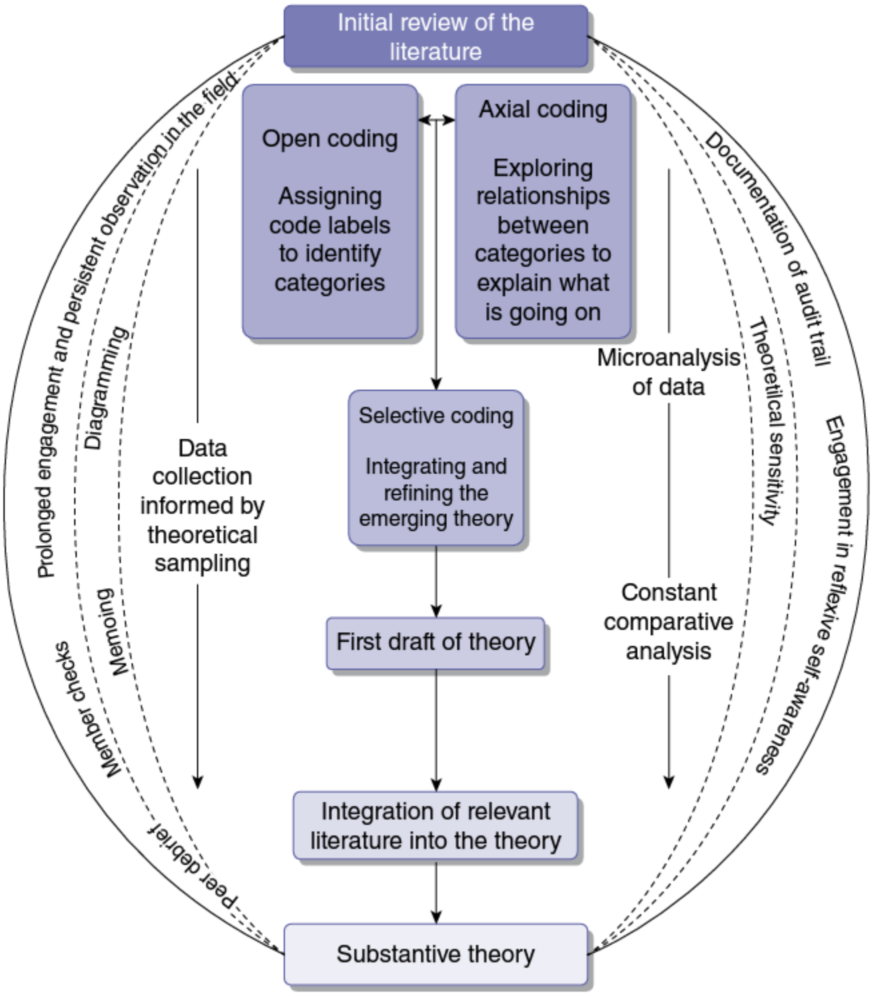

Initial review of the literature
Open coding
Assigning code labels to identify categories
Axial coding
Exploring relationships between categories to explain what is going on
Selective coding
Integrating and refining the emerging theory
First draft of theory
Integration of relevant literature into the theory
Substantive theory
Data collection informed by theoretical sampling
Microanalysis of data
Constant comparative analysis
Prolonged engagement and persistent observation in the field
Diagramming
Memoing
Member checks
Peer debrief
Documentation of audit trail
Theoretical sensitivity
Engagement in reflexive self-awareness

Figure 3-2. Schematic representation of the Grounded Theory approach (Corbin & Strauss, 2015, p.344)

Coding

Thirty-six audio-recorded interviews were transcribed by the researcher. Written style transcription conventions were deemed most suitable because the transcripts were not intended for linguistic or discourse analysis but “for reporting the subject’s accounts in a readable public story” (Kvale & Brinkmann, 2009, p.181). Therefore, pause length, intonation emphasis, or emotional expressions were not included. The transcripts were then checked for accuracy against the recordings and coded using the software ATLAS.ti.

The initial coding process (i.e. open coding) loosely followed the one described by Creswell (2014, p.268) and consisted of several steps: repeatedly reading through the data, dividing it into segments of information, and labelling them with codes. Fifty initial codes were developed.

For the axial coding stage, the codes were refined, and connections among the codes, as well as between the codes and other study data, were identified. In particular, the codes were streamlined to make them more focussed on the research questions. Some codes that did not directly feed into the research questions were discarded. Additionally, some new codes were generated, based on the comparison of the pre- and post-training interview data, as well as on the comparison of the interview data with other data from the study. For instance, a striking difference between the participants' use of specifications and example items in their pre- vs post-training item-writing only transpired after the pre- and post-training interview responses were compared; therefore, two new codes 'Use of example items' and 'Use of specifications' were created. Some code names were changed to make the coding scheme more comprehensive. For example, the analysis of quantitative data revealed different trends in evaluating items on objectively- and subjectively-scored criteria. To allow for comparison of quantitative and qualitative data, all criteria-related codes were sub-categorised into 'objective' (participants' discussions of objectively-scored criteria) and 'subjective'.

The resulting final version of the coding scheme included 41 codes in two categories: *Item-writing skill development* (28 codes), and *Role of the training* (13 codes) (see *Appendix 9*). After the final version of the codes was established, 10% of the interview data was double-coded by a fellow PhD student specialising in language testing from Lancaster University's Department of Linguistics and English Language. The overall coder agreement was 85%. Any coding differences were then discussed between the two coders and agreement was reached.

At the final stage of the analysis, findings from the analysis of codes in the *Item-writing skill development* category were used to answer RQ2, '*How did the participants' item-writing skill develop following the training, as perceived by the participants in interviews?*'. Findings from the *Role of the training* code category, together with findings from the analysis of feedback questionnaire responses, were used to answer RQ3, '*What role did the participants perceive the training played in their item-writing skill development?*'.

This section provided a detailed description of the Pretest-Posttest study, one of the two main studies of this research project. The second main study, the Course Feedback study, is described in the next section.

3.5.2 Main study 2: Course Feedback

This section provides a detailed account of the second element of the main study, that is a Course Feedback study aiming to gain insights into trainees' views on the role of the item-writing training course in developing their item-writing skills.

Literature in the fields of language teaching and human resource development provide similar guidelines and suggest similar methods for evaluating participant reactions to a training course. As both fields are relevant to the present research project, they are reviewed together in this section. Kirkpatrick and Kirkpatrick (2006) give some recommendations on how to collect feedback from training participants. In particular, they highlight the importance of being clear on what is to be evaluated, ensuring that reactions from all course participants are gathered and that the responses are honest. Phillips (1991) suggests the following feedback areas: "program content, instructional materials, out-of-class assignments, methods of presentation, instructor/speaker, program relevance, facilities, general evaluation, and planned improvement" (p.161). Although not all areas are relevant to the item-writing course researched in this study, the recommendations were taken into consideration when deciding on the content of the feedback questionnaires for this study, as outlined in Section 3.5.2.1.

The two main methods suggested in the literature for obtaining participants' feedback are questionnaires and interviews, with many researchers believing that "a combination of interviewing and questionnaires works best" (Lynch, 2003, p.130). Newby (1992) made a comparison of interview and questionnaire strengths: questionnaires can offer anonymity, are fast to administer, can gather responses from large samples, and allow for more straightforward data coding; interviews, on the other hand, "can yield better response levels, … allow for probing and follow-up questions" (p.79), and make checking understanding possible. Recommendations on interview design were reviewed in Section 3.5.1.5 while recommendations on questionnaire design are summarised below.

Both closed and open questions are favoured for inclusion in feedback questionnaires, but for different reasons. Weir and Roberts (1994) believe that closed questions generate data that is easy to analyse statistically and to cross-compare; however, closed questions are less informative and can lead to overlooking important opinions from participants. Open questions "can obtain richer, more divergent information that is not limited to the areas pre-determined by the evaluator" (Weir & Roberts, 1994, p.154). Weir and Roberts warn against including leading, ambiguous, over-general, offensive, presumptuous, hypothetical, and jargon questions that have the potential to contaminate study findings. Newby (1992) suggests starting a questionnaire with more general questions and then moving to "more specific questions on each particular theme and from the more familiar to less familiar" ones (p.82). He believes that questions about a particular aspect of the course should be grouped together in a sequence that "makes psychological sense to respondents" (p.82).

Questionnaire piloting is seen in the literature as crucial because there is often only one opportunity for the main data collection, while an unpiloted questionnaire can result in irrelevant or unpredictable responses. Two stages of piloting are recommended: 1) with several colleagues who are experts in the field, and 2) with a small sample of respondents from the target population. Newby (1992) suggests interviewing colleague respondents while they are answering the questionnaire, to find out whether they understand each question and whether it is difficult to answer. Responses from the second pilot should be subjected to "a dummy analysis" (Weir & Roberts, 1994, p.158), and any questions that did not generate useable data should be discarded or rewritten.

As concerns the analysis of collected responses, two approaches are suggested: quantitative analysis of closed question responses and qualitative analysis of responses to open questions. Phillips (1991) recommends using "the simplest statistics possible… to draw the proper conclusions with the data" (p.193) and to avoid over-complications. Frequency distributions, measures of central tendency and dispersion are seen as sufficient to analyse yes/no and Likert scale-style responses. For analysing responses to open-ended questions, Newby (1992) suggests the following procedure: 1) reviewing the raw data, 2) finding key words or phrases to summarise each response, 3) establishing response categories, and 4) analysing responses in each category in accordance with research questions. Weir and Roberts (1994) also warn that categorising responses can be somewhat subjective and call for triangulation "through the use of different methodological procedures in studying the same programme" (p.160).

In the present study, participants' views on the item-writing course were gauged in two manners: 1) through four feedback questionnaires administered at different times of the course, and eliciting quantitative and qualitative data; and 2) through feedback elicited during the post-training interviews, that is qualitative data from that part of the post-training interviews where participants were asked to provide feedback on the training. This combination of questionnaire and interview methods reflects the recommendations from the literature (Lynch, 2003; Weir & Roberts, 1994). Below, the data collection and analyses methods of this Course Feedback study are described.

3.5.2.1 Data collection

Feedback questionnaires

Four feedback questionnaires (Appendices 10-13) were administered to participants throughout the course to ensure feedback continuity, as recommended by Phillips (1991). Participants were asked to respond to a questionnaire after every two modules of the course, as well as upon course completion. Feedback Questionnaires (FQ) 1 to 3 addressed specific areas for evaluation: course materials (FQ1), course activities (FQ2), course structure (FQ3), and use of technology (FQ3). The final FQ repeated the main questions from the three preceding questionnaires to detect any changes in participants' attitudes over time.

Both closed and open questions were used: closed questions required Likert-scale responses, while open questions asked participants to justify or elaborate on the responses to closed questions. Each questionnaire, as recommended by Newby (1992), started with more general questions about the course aspect of interest, and proceeded to specific questions about the helpfulness for item-writing skills development of individual materials, activities, or technology.

All questions in the four questionnaires were designed to elicit data in response to RQ3 of this study: 'What role did the participants perceive the training played in their item-writing skill development?' The specific questions in each questionnaire were formulated to make an explicit connection between a particular material, activity or piece of technology and participants' item-writing skill development. For example, question 9 of FQ1 asked participants to "indicate how useful you feel the following materials from modules 1 and 2 were **in helping to develop your item-writing skills**" *(Appendix 10)* and to explain their choice of response. General questions, although not explicitly worded in terms of item-writing skill development, were also meant to contribute to answering RQ3 of this study. In these questions, participants were asked about the course materials' usefulness, interest, user-friendliness, and quality; about the course activities' usefulness, interest, and user-friendliness; about the course structure's clarity, flexibility, and pace; and about the course technology's usefulness, supportiveness, and user-friendliness. All these course qualities were felt to be related to the development of item-writing skill during the course, as explained below.

Studies in the field of educational psychology provide empirical proof that interest is connected to student motivation - a positive relationship has been found between measures of interests and measures of intrinsic motivation (Weber, 2009; Frymier et al., 1996). The positive relationship was explained using the notion of 'self-intentionality', whereby "interest-related goal is compatible with one's preferred values and ideals of the growing self" (Krapp, 1999, p.26). Research also suggests that interest is correlated with the notion of self-efficacy: interested engagement is often accompanied by increased self-efficacy and leads to improved performance (Renninger & Hidi, 2016).

Questions about the user-friendliness, quality and clarity of course materials/technology were included in the interview schedule in the belief that a course's user-friendliness and quality have an effect on learning outcomes. Materials/technology that are not user-friendly and/or are of low-quality might provoke negative feelings among learners and decrease their motivation for learning. On the other hand, user-friendly high-quality materials/technology might contribute to positive feelings towards the learning process. A positive attitude is considered one of the affective factors that directly contributes to motivation and, consequently, to improving learning outcomes (Dornyei, 1990).

To account for the nature of the training course researched in this study, participants were also asked about the pace of the course and the flexibility of the course structure. All participants were working full-time while doing the training; therefore, it was important to investigate how well the course was able to accommodate participants' busy work schedules. It was felt that a course compatible with participants' other responsibilities might result in better learning (i.e. item-writing skill development), while a course that is in conflict with participants' duties in other areas of their life might lead to frustration, low learning rates, and course drop-outs.

To give an illustration of what each feedback questionnaire contained, a brief outline of FQ1 (*Appendix 10*), conducted after Module 2 of the course, is provided below. This questionnaire focussed on the course materials. It started with four Likert-scale questions about the course materials' (1) usefulness, (2) interest, (3) user-friendliness, and (4) quality. After each closed question, participants were required to elaborate on their answers in a comment box. The second part of the questionnaire asked for participants' feedback on individual materials in Modules 1-2. Respondents were asked to evaluate each piece of material on a 6-point scale for its usefulness in item-writing skills development, and then to elaborate on their responses. Finally, in an open question, participants were asked to provide any further suggestions on the improvement of the materials.

The questionnaires were hosted on the Qualtrics online survey management platform and underwent two stages of piloting. First, each questionnaire was pre-piloted with a fellow PhD student from the Department of Linguistics and English Language, Lancaster University, who is an expert in language training. She was asked to comment on clarity and usefulness of each question. Following her suggestions, some questions were reworded to make them clearer, and then piloted with Cohort 2 of the item-writing course. 'Dummy analysis' (Weir & Roberts, 1994, p.158) was run to determine whether the questionnaires yielded useable data. Only minor changes to the questions had to be made. For example, for a FQ2 question that asked participants to comment on individual activities, the activities were rearranged in the order they were performed during the course, rather than by activity type as they had been presented initially, to better stimulate participants' recall. A FQ3 question asking about the use of interactive activities was made clearer by specifying the activities. Further changes became necessary when the course was updated for Cohort 3 run (see Section 3.2 for a discussion of how the course evolved throughout the three cohorts). In particular, those questions that focussed on individual materials/activities, had to be updated.

The questionnaires, in their final version, were administered during Cohort 3. To guarantee honesty of opinions, participants were offered full anonymity. One week was allowed for completion of each FQ, after which a general reminder was sent to encourage participants to provide their responses, if they hadn't already done so. FQ1 was answered by 19 participants, FQ2 – 22 participants, FQ3 – 21 participants, and the Final FQ - 19 participants.

Post-training interviews

The Course Feedback study also drew on the retrospective semi-structured interviews conducted with Cohort 3 participants after they had completed the post-training item-writing assignment. The interview data collection was described in Section 3.5.1.5. Specifically, for the Course Feedback study, the following questions were included in the interview schedule:

- Please tell me more about the item-writing course you took. Do you think the course has helped you in any way to write items?
- Is there anything else you would like to tell me about the item-writing course?

Follow-up questions were used to prompt interviewee responses:

- Were there any particular aspects of the training course that have helped you in writing the items?
- Is there anything particular not covered in the course and which would have helped you to write the items?

If necessary, interviewees were further prompted to talk about particular aspects of the course, such as course materials and activities. However, if participants were forthcoming about their training experience, they were allowed to discuss the issues they wanted to focus on. The theoretical basis for the questions, as well as their connection to RQ3 of this study, are discussed in Section 3.5.1.5.

3.5.2.2 Data analyses

This section describes the methods used to analyse the feedback questionnaire and interview data.

Feedback questionnaires

Descriptive statistics were obtained for all quantitative responses: mean, range, and frequency distributions. Responses to FQ1-3 and the Final FQ were compared to detect any changes in participants' attitudes over time. To allow for valid comparisons, percentages were obtained because the number of respondents varied for individual questionnaires. Findings from questionnaires were summarised in tables and charts.

For open-ended questions, key themes were identified through multiple readings. Summaries of responses to individual questions were produced and, where relevant, comparisons made between the FQ1-3 and Final FQ findings.

Post-training interviews

The interview responses to course-related questions were coded and analysed together with other interview responses, as described in Section 3.5.1.5. Thirteen feedback-related codes were identified and combined into the category *Role of the training*. The category contained four sub-categories: *Training materials* (5 codes), *Training activities* (5 codes), *Course structure* (1 code), and *Use of technology* (2 codes). The sub-categories are identical to the four feedback areas in the FQs. In the final stage of analysis, the FQ findings and course-related interview findings were brought together to answer RQ3 '*What role did the participants perceive the training played in their item-writing skill development?*'.

3.6 Chapter summary

This chapter described the research design and methods used in this study. First, the chapter described the background to the research and the overall research design. Second, the pilot study was overviewed. The chapter then proceeded to discuss in detail two main studies of the project: the Pretest-Posttest study and the Course Feedback study. A description of the Pretest-Posttest study included information about the item-writing assignment and item-writer trainees. Items produced by the trainees for the pre- and post-training assignments were evaluated by expert judges against a rating scale. Various statistical measures were applied to the evaluations to determine 1) judges' agreement, 2) changes in item quality pre- to post-training, and 3) individual item-writer variation. The interviews were conducted with willing participants upon completion of each assignment and analysed using the Grounded Theory approach. The Course Feedback study examined participants' reactions to the course through feedback questionnaires and post-training interview questions.

The chapter that follows presents the main study results organised into three sections according to the three research questions of this project.

Chapter 4 Results

4.1 Introduction

This chapter presents the study’s findings, organised according to the three research questions:

1. How did the quality of items produced by novice item writers change from before to after an online item-writing training course?
2. How did the participants’ item-writing skills develop through the training, as perceived by the participants in interviews?
3. What role did the participants perceive the training played in their item-writing skill development?

Section 4.2 describes the quantitative findings related to the first research question. It is divided into four sub-sections: Section 4.2.1 presents findings from the descriptive statistics; Section 4.2.2 discusses findings from the comparative statistical analyses, Section 4.2.3 presents findings from the gain ratio statistics, while Section 4.2.4 provides an integrated summary of the quantitative findings. Section 4.3 reports on qualitative findings related to the second research question, while Section 4.4 describes findings related to the third research question, based on two types of data: feedback questionnaires (4.4.1) and post-course interviews in that part where participants discussed the course they had completed (4.4.2).

4.2 Item quality pre- vs. post-training (RQ1)

This section is related to the first research question, “*How did the quality of items produced by novice item writers change from before to after an online item-writing training course?*”, and reports on the findings from the quantitative item evaluations. To examine changes in item quality from before to after the training, descriptive statistics for the pre- and post-training item evaluations were analysed contrastively (4.2.1). The significance of changes in scores was tested by means of Wilcoxon signed-rank tests (4.2.2). Additionally, gain ratios were calculated to explore individual item-writer variations (4.2.3). A summary of the quantitative findings is presented in Section 4.2.4.

4.2.1 Findings from the descriptive statistics

As described in the Methodology Chapter (Section 3.5.1.4), the item evaluations are comprised of scores on objectively-scored criteria calculated by the researcher, and judgements made on subjectively-scored criteria by three reviewers working independently. For each participant's item, the final score for each subjectively-scored criterion was arrived at by using the median of the three independent judgments on that criterion. Descriptive statistics of item evaluations were obtained and interpreted separately for each item type: A2 and C1 grammar items (4.2.1.1), B2 writing prompts (4.2.1.2), and B1 listening tasks (4.2.1.3).

4.2.1.1 Findings on the A2 and C1 grammar items

Each of the 25 participants produced one A2 and one C1 multiple-choice grammar item both for the pre-training and the post-training assignments. The items were evaluated on 19 criteria: ten objectively-scored, eight subjectively-scored (*Appendix 6*) and an overall item acceptability score. The evaluation scale for each criterion spanned through three bands from '0' to '2' (see Section 3.5.1.4 of the Methodology Chapter for more detail). Descriptive statistics were obtained for: a) the total sum of scores on all criteria together for each item, b) the sum of scores separately on the objectively-scored and subjectively-scored criteria for each item, and c) the scores on each individual criterion.

Pre- and post-training total item scores and the overall item acceptability score

Descriptive statistics for the total item scores are presented in *Table* .

Table 4-1. Descriptive statistics for the total scores for A2 and C1 grammar items

	Range	Min	Max	Mean		SD	Skewness		Kurtosis	
				Statistic	SE		Statistic	SE	Statistic	SE
Total score for A2 grammar items										
Pre-training	10	25	35	**30.56**	.62	3.12	-.29	.46	-1.14	.90
Post-training	12	24	36	**33.16**	.47	2.37	-2.52	.46	8.93	.90
Total score for C1 grammar items										
Pre-training	17	19	36	**30.36**	.73	3.67	-1.05	.46	2.57	.90
Post-training	5	30	35	**33.08**	.34	1.68	-.48	.46	-.78	.90

Pre-training, out of a maximum possible score of 36, no grammar item scored lower than 19, with the score range for C1 items being much wider than that for A2 items (17 compared to

10, respectively). However, the much wider C1 range is explained with just one outlier (see *Figure 4-2*), indicating that one participant produced an item much weaker than all the other participants' items. No A2 item achieved the maximum possible total score pre-training, while two C1 items did. The skewness, although negative in both cases, displays different characteristics. For A2 items, the degree of skewness (obtained by dividing the skewness statistic by its standard error) is -.63 and is within the normal distribution parameters (Green, 2013, pp.44-45). For C1 items, the degree of skewness is -2.28, which is a substantial departure from symmetry. The density of distribution is also different for the two item types. There is more variability in the A2 item-quality scores, which is manifested with the platykurtic distribution (kurtosis is -1.14), while C1 scores have a leptokurtic distribution (kurtosis is 2.57) clustering at the higher end of the curve.

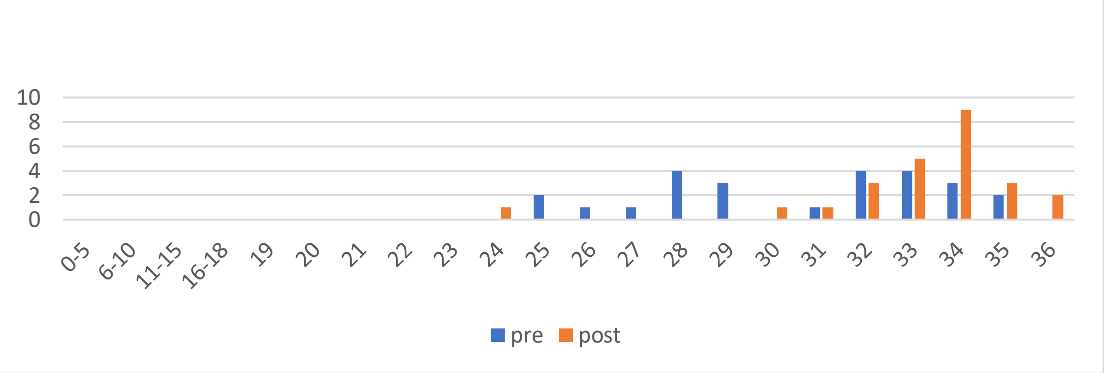


Figure 4-1. Descriptive statistics for the total scores, A2 grammar items

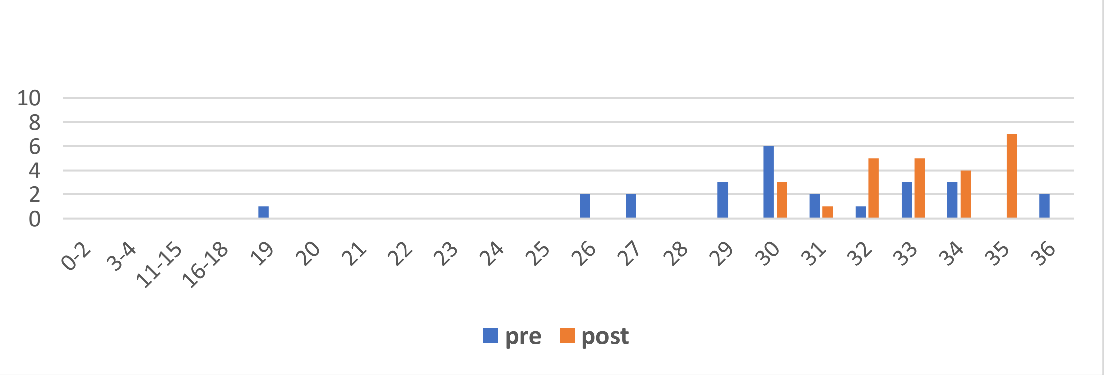


Figure 4-2. Descriptive statistics for the total scores, C1 grammar items

After the training, most participants' grammar items received higher total scores. Namely, pre-training, 14 participants produced A2 and 17 produced C1 items that scored 30 or higher, while everyone's C1 item and all-but-one's A2 item scored 30+ post-training. The A2 items' score range post-training was much wider than that for the C1 items but the wider range was due to only one outlier (*Figure 4-2*). Two A2 items achieved the highest possible score post-training, compared to none pre-training. However, while two C1 items scored the maximum of 36 pre-training, none did so after it. Moreover, the post-training total scores for A2 items were generally higher than those for C1 items, with the overall scores displaying a highly negatively skewed leptokurtic distribution (degree of skewness -5.47). This finding is supported by the scores on the overall acceptability criterion (*Table 4-2*): 20% more A2 items scored band '2' following the training, while 8% less C1 items did.

Table 4-2. Descriptive statistics for the overall item acceptability of A2 and C1 grammar items

	PRE-TRAINING					**POST-TRAINING**				
	Frequencies			**Mean**	**SD**	**Frequencies**			**Mean**	**SD**
	Band 0	**Band 1**	**Band 2**			**Band 0**	**Band 1**	**Band 2**		
A2	1	18	6	**1.20**	.50	1	13	11	**1.40**	.58
C1	3	13	9	**1.24**	.66	1	17	7	**1.24**	.52

These findings indicate that, although there was a larger proportion of participants whose grammar items scored quite high following the training - particularly for A2 items - not many participants managed to achieve the quality necessary for inclusion in a live test (Band 2), with 56% of A2 and 72% of C1 post-training items requiring further revision. Therefore, more detailed analysis of scores on individual criteria is necessary to identify which specification requirements posed more difficulty for these novice item writers. As the criteria can be divided into objectively-scored (by the researcher) and subjectively-scored (judged by reviewers), below, the discussion of the results is arranged by the criteria type.

Pre- and post-training total scores on the sum of the objectively-scored and of the subjectively-scored criteria

Before the training, the total scores for the objectively-scored criteria ranged between 12 and 20 (*Table 4-3*). A similar range was observed for the subjectively-scored criteria (*Table 4-4*).

Although the percentage of highest scoring items was generally low, it was lower for A2 items – 16% achieved the maximum total score on the objectively-scored (*Figure 4-3*) and 12% on subjectively-scored criteria (*Figure 4-5*) – compared to 24% (*Figure 4-4*) and 20% (*Figure 4-6*) respectively for C1 items.

Table 4-3. Descriptive statistics for the total scores on the objectively-scored criteria of A2 & C1 grammar items

	Range	**Min**	**Max**	**Mean**		**SD**	**Skewness**		**Kurtosis**	
				Statistic	**SE**		**Statistic**	**SE**	**Statistic**	**SE**
A2 grammar items										
Pre-training	8	12	20	**17.44**	.47	2.33	-1.01	.46	.04	.90
Post-training	6	14	20	**19.28**	.26	1.31	-2.99	.46	11.03	.90
C1 grammar items										
Pre-training	8	12	20	**16.76**	.51	2.57	-.19	.46	-1.02	.90
Post-training	4	16	20	**19.00**	.26	1.29	-1.26	.46	.59	.90

After the training, the minimum scores were much higher, especially for C1 items, which resulted in the total score range on both the objectively-scored and subjectively-scored criteria being narrower compared to pre-training. This finding demonstrates that the overall item quality was higher following the training. Post-training, total scores for the objectively-scored criteria of both the A2 and C1 items (*Figure 4-3; Figure 4-4*) clustered very closely at the higher end of the distribution. However, total scores for the subjectively-scored criteria (*Figure 4-5; Figure 4-6)* were still quite widely distributed. This finding indicates that the participants wrote much better items with regard to the objectively-scored criteria following the training, while an improvement in the quality on the subjectively-scored criteria was far less pronounced.

Table 4-4. Descriptive statistics for the total scores on the subjectively-scored criteria of A2 & C1 grammar items

	Range	**Min**	**Max**	**Mean**		**SD**	**Skewness**		**Kurtosis**	
				Statistic	**SE**		**Statistic**	**SE**	**Statistic**	**SE**
A2 grammar items										
Pre-training	8	8	16	**13.16**	.44	2.21	-.92	.46	.19	.90
Post-training	6	10	16	**13.96**	.33	1.67	-1.21	.46	1.03	.90
C1 grammar items										
Pre-training	9	7	16	**13.60**	.43	2.16	-1.16	.46	2.02	.90
Post-training	4	12	16	**14.08**	.22	1.11	.03	.46	-.35	.90

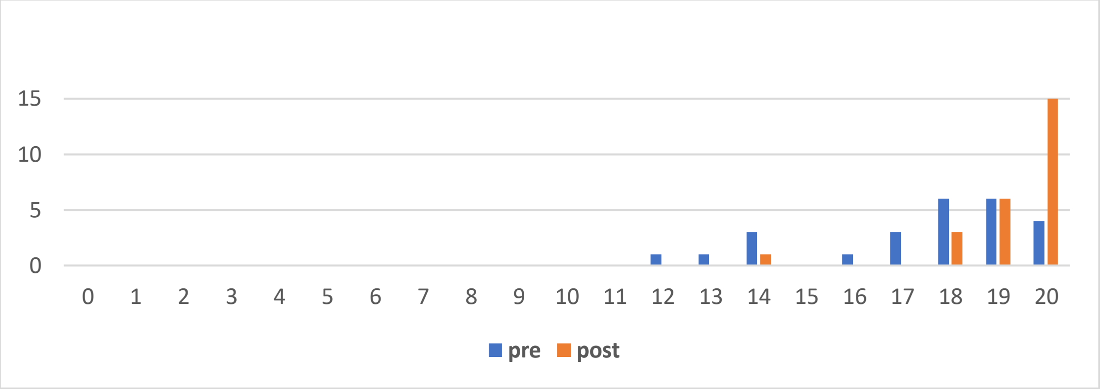


Figure 4-3. A2 grammar items' objectively-scored criteria: descriptive statistics for the total scores

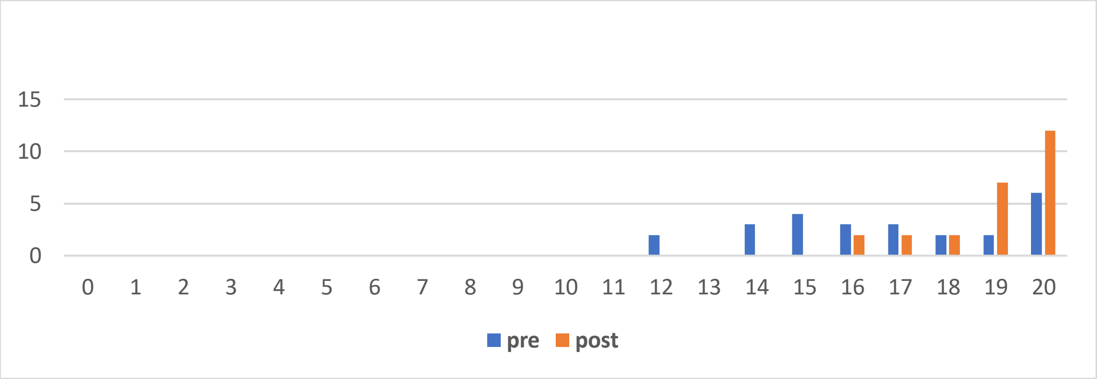


Figure 4-4. C1 grammar items' objectively-scored criteria: descriptive statistics for the total scores

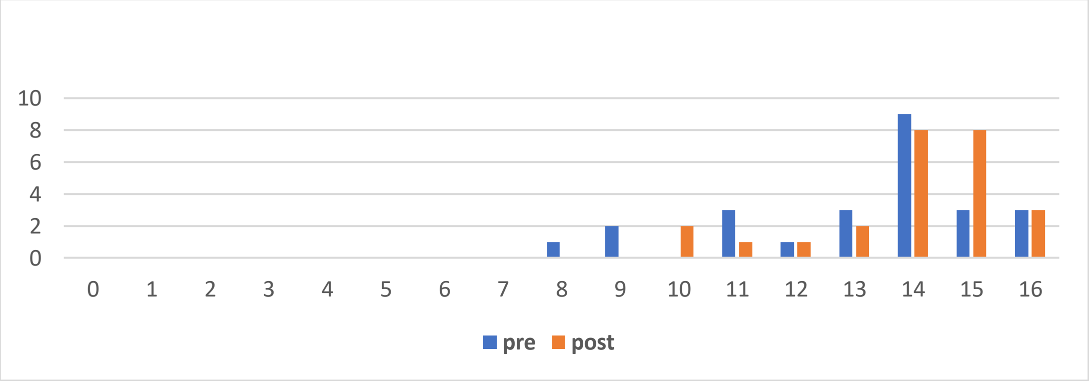


Figure 4-5. A2 grammar items' subjectively-scored criteria: descriptive statistics for the total scores

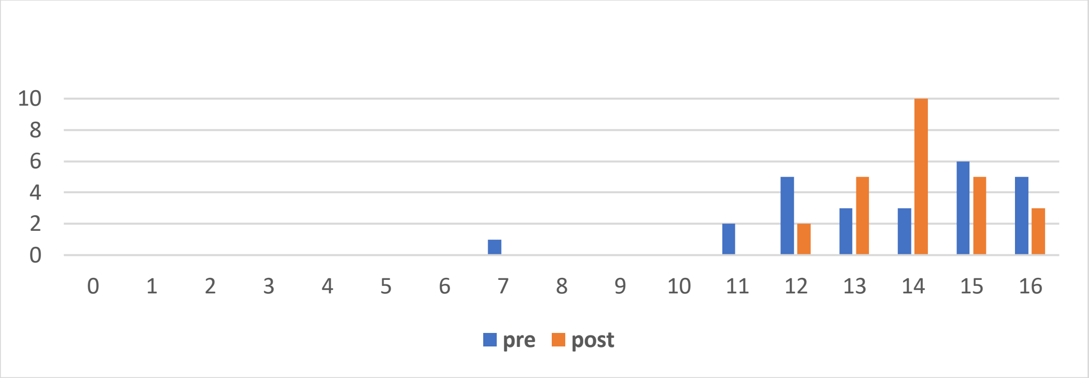


Figure 4-6. C1 grammar items' subjectively-scored criteria: descriptive statistics for the total scores

Pre- and post-training scores on individual objectively-scored criteria

The ten objectively-scored criteria used to evaluate A2 and C1 grammar items are presented in *Table 4-5* (see also *Appendix 6* for the complete Item Evaluation Scales, including descriptors for each band score).

Table 4-5. Objectively-scored criteria to evaluate grammar items

G1	Stem: max. 10 (A2) / 15 (C1) words including the key
G2	Stem: contains one gap only
G3	Options: 3 including the key and distractors
G4	Options: 1-3 words
G5	Options: there are no words at the beginning or the end of all options which can be integrated into the stem
G6	Key: indicated with asterisk
G7	Lexis: K1 (A2) / K1-5 (C1)
G8	Topic: appropriate at A2 / C1 level
G9	Function: appropriate at A2 / C1 level
G10	Spelling / grammar / punctuation of the stem and options: correct

Before the training (*Table 4-6*), most participants managed to write items that met the word-limit (G1, G4), item format (G2, G3), and vocabulary frequency (G7) criteria (M=1.84 to 2.0). The requirements that posed more difficulty involved formulating concise options (G5), choosing an appropriate topic and function (G8, G9), as well as indicating the key (G6) and proofreading the item (G10). The mean values for these criteria ranged from 1.2 to 1.68, with substantially more band '0' scores. While the A2 and C1 grammar items demonstrated similar trends, there was one difference: participants found it considerably more difficult to formulate concise options for C1 items (M=1.2) compared to A2 items (M=1.6).

After the training, mean values for the objectively scored criteria ranged between 1.72 and 2.0, which is much higher than pre-training. The criteria which scored high pre-training scored similarly high or higher post-training. Additionally, all the criteria with which the participants had problems pre-training had higher mean values following the training. This is because of much fewer band '0' scores (by 8.8% for A2 items and 10.4% for C1 items) and more band '2' scores (by 9.6% for A2 and 12% for C1 items). At the same time, the number of band '1' scores stayed almost the same (see *Appendix 14*).

Table 4-6. Descriptive statistics for the objectively-scored criteria of A2 and C1 grammar items

Criteria	PRE-TRAINING					POST-TRAINING				
	Frequencies			Mean	SD	Frequencies			Mean	SD
	Band 0	Band 1	Band 2			Band 0	Band 1	Band 2		
Objectively-scored criteria: A2 grammar items										
G1	0	1	24	**1.96**	.20	0	0	25	**2.00**	.00
G2	1	0	24	**1.92**	.40	0	2	23	**1.92**	.28
G3	0	0	25	**2.00**	.00	0	0	25	**2.00**	.00
G4	1	0	24	**1.92**	.40	1	0	24	**1.92**	.40
G5	2	6	17	**1.60**	.64	0	1	24	**1.96**	.20
G6	6	0	19	**1.52**	.87	0	0	25	**2.00**	.00
G7	1	2	22	**1.84**	.47	0	3	22	**1.88**	.33
G8	4	0	21	**1.68**	.75	1	1	23	**1.88**	.44
G9	6	1	18	**1.48**	.87	2	2	21	**1.76**	.60
G10	5	2	18	**1.52**	.82	0	1	24	**1.96**	.20
Objectively-scored criteria: C1 grammar items										
G1	1	1	23	**1.88**	.44	1	1	23	**1.88**	.44
G2	1	1	23	**1.88**	.44	0	1	24	**1.96**	.20
G3	0	0	25	**2.00**	.00	0	0	25	**2.00**	.00
G4	1	0	24	**1.92**	.40	0	0	25	**2.00**	.00
G5	7	6	12	**1.20**	.87	0	5	20	**1.80**	.41
G6	7	0	18	**1.44**	.92	1	0	24	**1.92**	.40
G7	0	2	23	**1.92**	.28	0	1	24	**1.96**	.20
G8	6	1	18	**1.48**	.87	2	1	22	**1.80**	.58
G9	4	2	19	**1.60**	.76	2	3	20	**1.72**	.61
G10	5	4	16	**1.44**	.82	0	1	24	**1.96**	.20

Pre- and post-training scores on individual subjectively-scored criteria

The eight subjectively-scored criteria used to evaluate A2 and C1 grammar items are presented in *Table 4-7* (see also *Appendix 6).*

Table 4-7. Subjectively-scored criteria to evaluate grammar items

G11	Stem: provides enough context to ensure that the intended construct is tested, including restricting the number of possible correct answers
G12	Distractors: strong, plausible
G13	Distractors: not grammatically correct within the stem
G14	Distractors: grammatically correct as a stand-alone
G15	Key: does not stand out from the distractors
G16	Grammar exponent: directly targeted in the item
G17	Grammar of the stem / key: 'standard' English, i.e. not dialect, jargon, etc.
G18	Content: appropriate, culturally unbiased, not disturbing, suitable for a general-purpose test (i.e. not a specific purpose test)

As evident from *Table 4-8,* three subjectively-scored criteria – content fairness (G18), use of standard English (G17), and distractors not being grammatically correct within the stem (G13) – were not difficult for most item writers to meet pre-training. Four criteria proved more difficult for the untrained participants to conform to: the construct-related ones (G11, G16), and two option-related ones (G14, G15), with the mean scores ranging between 1.48 and 1.76. A2 items scored slightly lower on the option-related criteria, and C1 items on one construct-related criterion. By far the lowest were the scores awarded for distractor strength and plausibility (G12), with A2 items (M=0.92) scoring substantially lower than C1 ones (M=1.2) due to a larger number of '0' scores for A2 items. Overall, construct- and distractor-related criteria seemed more challenging than other subjectively-scored criteria before the training. Among these, the participants found writing strong plausible distractors most difficult, especially when creating A2 items.

Table 4-8. Descriptive statistics for the subjectively-scored criteria of A2 and C1 grammar items

Criteria	PRE-TRAINING					POST-TRAINING				
	Frequencies			Mean	SD	Frequencies			Mean	SD
	Band 0	Band 1	Band 2			Band 0	Band 1	Band 2		
Subjectively-scored criteria: A2 grammar items										
G11	3	5	17	**1.56**	.71	1	12	12	**1.44**	.58
G12	8	11	6	**.92**	.76	3	10	12	**1.36**	.70
G13	0	2	23	**1.92**	.28	0	0	25	**2.00**	.00
G14	2	7	16	**1.56**	.65	0	1	24	**1.96**	.20
G15	2	6	17	**1.60**	.64	2	6	17	**1.60**	.64
G16	1	7	17	**1.64**	.57	1	8	16	**1.60**	.58
G17	0	0	25	**2.00**	.00	0	0	25	**2.00**	.00
G18	0	0	25	**2.00**	.00	0	0	25	**2.00**	.00
Subjectively-scored criteria: C1 grammar items										
G11	2	9	14	**1.48**	.65	0	7	18	**1.72**	.46
G12	5	10	10	**1.20**	.76	3	17	5	**1.08**	.57
G13	0	0	25	**2.00**	.00	0	1	24	**1.96**	.20
G14	1	4	20	**1.76**	.52	0	0	25	**2.00**	.00
G15	0	6	19	**1.76**	.44	0	10	15	**1.60**	.50
G16	2	5	18	**1.64**	.64	1	2	22	**1.84**	.47
G17	0	1	24	**1.96**	.20	0	0	25	**2.00**	.00
G18	0	4	21	**1.84**	.37	0	3	22	**1.88**	.33

Post-training, the range of mean values on the subjectively-scored criteria was still wide: 1.36-2.0 for A2 and 1.08-2.0 for C1 items. This indicates that some criteria continued to pose considerable difficulty for item writers after the training. All criteria that scored high pre-training scored similarly high or higher post-training. While similar to the objectively-scored criteria the number of band '0' scores was lower and band '2' scores higher following the training, the difference was less pronounced: band '0' scores were 4.5% fewer for A2 and by 3% for C1 items, while there were 5% and 4% more band '2' scores, respectively (see *Appendix 14*).

Post-training, the lowest mean scores were for distractor strength and plausibility (G12, A2 M=1.36, C1 M=1.08). This requirement seems to have posed the greatest difficulty to the participants – for both A2 and C1 items, both before and after the training. However, while the A2 mean score on this criterion was much higher following the training (1.36 compared to 0.92 pre-training), for C1 items it was, in fact, lower (1.08 compared to 1.20 pre-training). Lower mean values after the training is an unexpected result that was not detected for any of the objectively-scored criteria. However, for the subjectively-scored criteria this is not unique.

For A2 items, mean values for two construct-related criteria (G11 and G16) were somewhat lower after the training, while C1 mean values on the same criteria were substantially higher. At the same time, three C1 mean scores were lower post-training compared to pre-training, all three being distractor-related (G12, G13, and G15). However, A2 mean values on two of these criteria (G12 and G13) were higher post-training. Overall, A2 and C1 grammar items seemed to display opposing trends regarding the subjectively-scored criteria identified as most challenging prior to the training. Following the training, participants targeted the intended construct much better in C1 items but worse in A2 items. On the other hand, participants generally wrote better A2 distractors, while the quality of C1 distractors was weaker.

4.2.1.2 Findings on the B2 writing prompts

The writing prompts were evaluated on 17 criteria: five objectively-scored, 11 subjectively-scored (*Appendix 6*) and an overall item acceptability score, using a three-band scale. The descriptive statistics were obtained in the same way as for the grammar items.

Pre- and post-training total item scores and the overall item acceptability score

Pre-training, 22 participants' writing prompts received a total score of 27 or higher out of a maximum possible score of 32, while prompts from three participants were substantially lower quality than the rest (*Figure 4-7).* This means that, even before the training, most participants were able to produce reasonably good-quality writing prompts. At the same time, only three participants' prompts received the maximum total score pre-training.

Table 4-9. Descriptive statistics for the total scores of the B2 writing prompts

	Range	Min	Max	Mean		SD	Skewness		Kurtosis	
				Statistic	SE		Statistic	SE	Statistic	SE
Pre-training	15	17	32	**28.64**	.65	3.26	-2.29	.46	6.47	.90
Post-training	6	26	32	**29.76**	.34	1.71	-.83	.46	.12	.90

The finding is supported by the overall acceptability statistics (*Table 4-10*) – only eight prompts were given 'the green light' by the reviewers (M= 1.24). In other words, although many item writers produced reasonably solid drafts, very few of those items were fully ready for live testing without further revision.

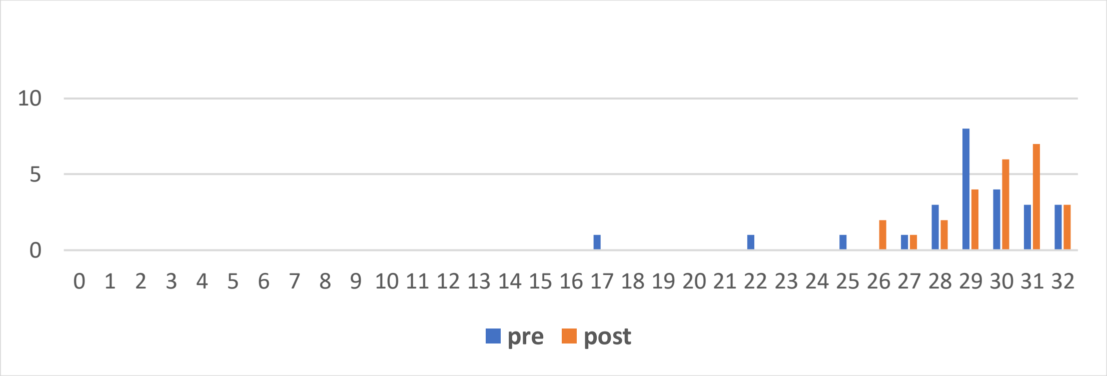


Figure 4-7. Descriptive statistics for the total scores, B2 writing prompts

Unlike pre-training, there were no outliers after the training (*Figure 4-7*), which is reflected in a much narrower total score range – from 15 pre-training to only 6 post-training (*Table 4-9*). The overall acceptability scores (*Table 4-10*) were similar to the ones pre-training, with one fewer prompt scoring band '0' and eight prompts scoring band '2' on each occasion. The statistics might indicate that, with respect to developing writing prompts, the training was most beneficial for the weakest participants; however, the training was insufficient for the participant cohort to start producing high-quality writing prompts that are immediately acceptable for live testing.

Table 4-10. Descriptive statistics for the overall item acceptability of B2 writing prompts

PRE-TRAINING					**POST-TRAINING**				
Frequencies			**Mean**	**SD**	**Frequencies**			**Mean**	**SD**
Band 0	**Band 1**	**Band 2**			**Band 0**	**Band 1**	**Band 2**		
2	15	8	**1.24**	.60	1	16	8	**1.28**	.54

Pre- and post-training total scores on the sums of the objectively-scored and of the subjectively-scored criteria

Before the training, there were substantial differences in the total scores on the objectively-scored and subjectively-scored criteria (*Table 4-11*). The total scores for the objectively-scored criteria were within normal distribution parameters (degree of skewness=-1.62; *Figure 4-8*), while there was a very large deviation from the normal distribution for scores on the subjectively-scored criteria (degree of skewness=-6.71; *Figure 4-9*). This is largely due to several outlier items, which scored much lower than the rest. While 40% of the writing prompts obtained the maximum possible total score on the objectively-scored criteria, only 16% achieved the same for the subjectively-scored criteria. These findings demonstrate that subjectively-scored criteria requirements were generally more challenging for the participants to meet; besides, there was a greater variation in the participants' ability to meet the subjectively-scored criteria compared to the objectively-scored ones. Therefore, the outliers identified at the beginning of this section were due to participants' varied ability to conform to the subjectively-scored criteria requirements before the training, while the participants were much more homogeneous in their ability to meet the objectively-scored requirements.

Table 4-11. Descriptive statistics for the total scores on the objectively-scored and subjectively-scored criteria of B1 writing prompts

	Range	Min	Max	Mean		SD	Skewness		Kurtosis	
				Statistic	SE		Statistic	SE	Statistic	SE
Total scores on the objectively-scored criteria										
Pre-training	4	6	10	**8.76**	.26	1.3	-.75	.46	-.40	.90
Post-training	4	6	10	**9.16**	.22	1.10	-1.34	.46	1.42	.90
Total scores on the subjectively-scored criteria										
Pre-training	13	9	22	**19.88**	.53	2.67	-3.11	.46	11.78	.90
Post-training	4	18	22	**20.60**	.26	1.32	-.59	.46	-.81	.90

Post-training, the total score range for the objectively-scored criteria was identical to the one pre-training, with slightly more prompts achieving higher scores, including more prompts that gained the maximum possible score (13 compared to 10 pre-training). The total score range for the subjectively-scored criteria post-training was much narrower compared to that pre-training (4 versus 13, respectively) due to the disappearance of outliers. Moreover, the improvement in the quality of writing prompts regarding the subjectively-scored criteria can be seen in the fact that twice as many prompts gained the maximum total score on the sum of the subjectively-scored criteria after the training.

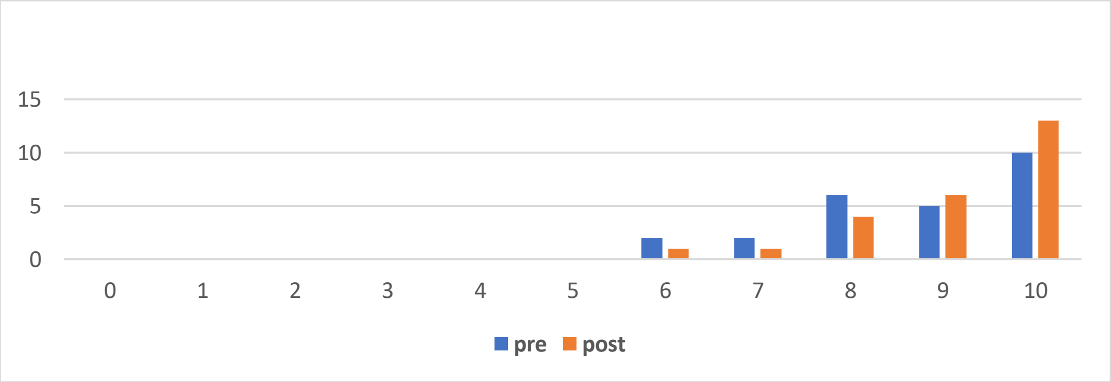


Figure 4-8. B2 writing prompts objectively-scored criteria: descriptive statistics for the total scores

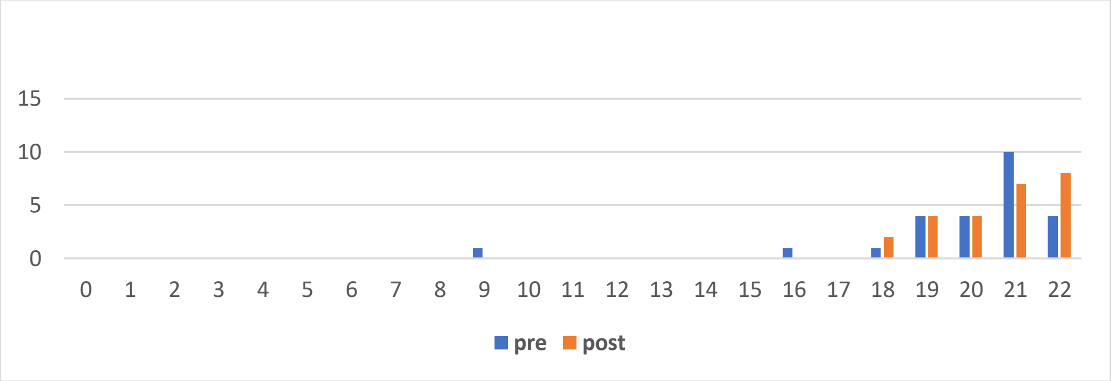


Figure 4-9. B2 writing prompts subjectively-scored criteria: descriptive statistics for the total scores

Pre- and post-training scores on individual objectively-scored criteria

The five objectively-scored criteria used to evaluate B2 writing prompts are presented in *Table 4-12* (see also *Appendix 6).*

Table 4-12. Objectively-scored criteria to evaluate B2 writing prompts

W1	Input message: 40-60 words
W2	Overall length of the prompt: 80-120 words
W3	Grammar: A1 – B1
W4	Lexis: K1 - K4
W5	Spelling / grammar / punctuation: correct

Before the training, the mean values for the writing prompts on the objectively-scored criteria (*Table 4-13*) ranged from 1.6 to 1.96, with very few band '0' scores awarded on any criterion. The participants were successful at meeting the prompt's word-limit (W2), grammatical range (W3), and vocabulary frequency (W4) requirements, which had also been the case for the grammar items. Interestingly, item writers found it more challenging to conform to the word-limit for the input message (W1, M=1.68) than the whole of the prompt (W2, M=1.96). The proofreading requirement (W5) received the lowest scores, which is again similar to what was found for the grammar items.

While most post-training mean values for the objectively-scored criteria of the grammar items were substantially higher compared to the pre-training ones, the pre- vs. post-training writing prompt mean values showed varying trends: the grammatical range, vocabulary frequency, and proofreading requirements (W3-W5) had higher post-training mean values, the mean value for the whole prompts' word limit (W2) stayed the same, while the mean value for the input messages' word limit (W1) was slightly lower. Notably, the unexpected post-training decrease in scores on the input messages' word-limit (W1) could have influenced the scores on the whole prompts' word limit (W2) because an input message is part of the prompt.

Table 4-13. Descriptive statistics for the objectively-scored criteria of B2 writing prompts

Criteria	PRE-TRAINING					POST-TRAINING				
	Frequencies			Mean	SD	Frequencies			Mean	SD
	Band 0	Band 1	Band 2			Band 0	Band 1	Band 2		
W1	1	6	18	**1.68**	.56	2	5	18	**1.64**	.64
W2	0	1	24	**1.96**	.20	0	1	24	**1.96**	.20
W3	0	4	21	**1.84**	.37	0	2	23	**1.92**	.28
W4	1	6	18	**1.68**	.56	1	2	22	**1.84**	.47
W5	3	4	18	**1.60**	.71	0	5	20	**1.80**	.41

Post-training, the band '2' score count was 6.4% higher across the five objectively-scored criteria (see *Appendix 14*). This was paralleled with a reduction in the number of band '1' (by 4.8%) and band '0' scores (by 1.6%). The small reduction in band '0' scores is unsurprising given their small number before the training and thus limited scope for further reduction.

Pre- and post-training scores on individual subjectively-scored criteria

The 11 subjectively-scored criteria used to evaluate B2 writing prompts are presented in *Table 4-14* (see also *Appendix 6).*

Table 4-14. Subjectively-scored criteria to evaluate B2 writing prompts

W6	Input message: a formal email / public notice
W7	Input message: clear and unambiguous
W8	Input message: suitable for testing, i.e. NOT a parody, not silly, humorous, sarcastic, etc.
W9	Input message: presents a plausible problem / issue / offer / opportunity which the candidate is expected to discuss
W10	Instruction: specifies the intended reader of the response email
W11	Instruction: specifies the purpose of the response email: complaining, suggesting alternatives, offering advice.
W12	Instruction: the purpose of the response email is plausible, i.e. the test-taker is asked to write a response for a plausible reason
W13	Instruction: the purpose of the response email is not too general and does not allow so much freedom to candidates as to result in vastly different responses
W14	Instruction: clear and unambiguous, not too wordy or excessive; includes the following information: "Write 120-150 words. You have 20 minutes."
W15	Intended response: the task encourages an original response and NOT copying from the input message
W16	Prompt (instructions + input message) content: appropriate, culturally unbiased, not disturbing, suitable for a general-purpose test (i.e. not a specific purpose test)

Pre-training, participants were most successful at producing an input message suitable for use in a test (W8), specifying the purpose of the response email in instructions (W11), and encouraging an original response from test-takers (W15) – the mean values for all three criteria equalled 1.92. At the same time, participants struggled with the input message's genre (W6, M=1.72), input message's plausibility (W9, M=1.60), and the clarity of the instruction (W14, M=1.60). The lower mean values are due to a large number of prompts scoring band '1' on these criteria, and few band '0' scores (*Table 4-15*).

Table 4-15. Descriptive statistics for the subjectively-scored criteria of B2 writing prompts

Criteria	PRE-TRAINING					POST-TRAINING				
	Frequencies			Mean	SD	Frequencies			Mean	SD
	Band 0	Band 1	Band 2			Band 0	Band 1	Band 2		
W6	1	5	19	**1.72**	.54	2	3	20	**1.72**	.614
W7	1	1	23	**1.88**	.44	0	1	24	**1.96**	.200
W8	1	0	24	**1.92**	.40	0	1	24	**1.96**	.200
W9	0	10	15	**1.60**	.50	0	8	17	**1.68**	.476
W10	0	4	21	**1.84**	.37	0	2	23	**1.92**	.277
W11	0	2	23	**1.92**	.28	1	0	24	**1.92**	.400
W12	1	2	22	**1.84**	.47	0	0	25	**2.00**	.000
W13	0	5	20	**1.80**	.41	0	2	23	**1.92**	.277
W14	2	6	17	**1.60**	.64	2	4	19	**1.68**	.627
W15	0	2	23	**1.92**	.28	0	3	22	**1.88**	.332
W16	1	2	22	**1.84**	.47	0	1	24	**1.96**	.200

After the training, the participants produced higher-quality prompts with regard to eight subjectively-scored criteria (*Table 4-15*). However, the post-training mean values on those criteria were not substantially higher because the pre-training mean values were already quite high. At the same time, the mean value for the requirement that the task should encourage an original response (W15) was slightly lower. The mean values for the input message genre (W6) and construct (W11) requirements stayed the same at 1.72 and 1.92, respectively.

Trends for the score frequency statistics for the subjectively-scored criteria were similar to those for the objectively-scored ones: because very few band ‘0’ scores were awarded pre-training (2.5% of the total score count), there were only 0.7% fewer band ‘0’ scores awarded post-training. The expert judges also awarded fewer band ‘1’ scores (5.2% fewer) but more band ‘2’ scores (5.9% more) after the training (see *Appendix 14*).

4.2.1.3 Findings on the B1 listening tasks

Listening tasks were evaluated on 27 criteria: 12 objectively-scored, 14 subjectively-scored (*Appendix 6*) and an overall item acceptability score, using a three-band scale.

Pre- and post-training total item scores and the overall item acceptability score

The pre-training total score statistics (*Table 4-16*) suggest that developing listening tasks posed more difficulty to untrained item writers compared to the other two item types.

Table 4-16. Descriptive statistics for the total scores of B1 listening tasks

	Range	Min	Max	Mean		SD	Skewness		Kurtosis	
				Statistic	SE		Statistic	SE	Statistic	SE
Pre-training	16	34	50	**42.04**	.83	4.17	-.50	.46	-.22	.90
Post-training	11	39	50	**45.24**	.62	3.10	-.21	.46	-.95	.90

The scores are almost equally distributed on each side of the mean (*Figure 4-10*), with the degree of skewness -1.08. No participant produced a listening task that scored the maximum, only one task scored 50 out of a maximum of 52, while most total scores clustered between 40 and 46. There was also a larger number of very low-quality tasks – so-called 'outliers' – compared to what had been the case for the other item types. The overall acceptability scores (*Table 4-17*) support these findings: seven tasks (28%) were rejected by the reviewers, with only three tasks (12%) considered acceptable for live testing without revision. These figures are much lower compared to those for the grammar and writing items where 25% to 35% of tasks respectively scored band '2' on the overall acceptability pre-training, with very few items rejected.

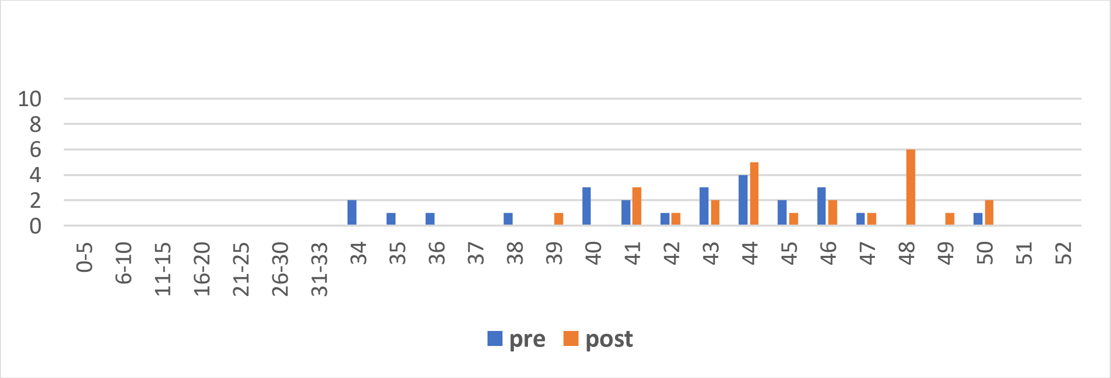


Figure 4-10. Descriptive statistics for the total scores, B1 listening tasks

The total score range decreased from 16 to 11 following the training (*Figure 4-10*) with nine participants having produced tasks that scored 48 or higher, compared to only one task scoring this high pre-training. At the same time, the maximum total score stayed at 50 with no task scoring the maximum possible. Moreover, the scores on the overall item acceptability were almost identical to the ones pre-training: two fewer tasks received band '0' but only three tasks scored '2', with most tasks still requiring revisions before they could be accepted for live testing.

Table 4-17. Descriptive statistics for the overall item acceptability of B1 listening tasks

PRE-TRAINING					POST-TRAINING				
Frequencies			Mean	SD	Frequencies			Mean	SD
Band 0	Band 1	Band 2			Band 0	Band 1	Band 2		
7	15	3	**.84**	.62	5	17	3	**.92**	.57

As well as being most challenging for participants pre-training, the quality of listening tasks produced saw the least overall improvement following the training out of all three task types. More detailed analyses of scores on individual criteria might clarify the reasons for this finding.

Pre- and post-training total scores on the sums of the objectively-scored and of the subjectively-scored criteria

Unlike the grammar and writing items, none of the listening tasks achieved the maximum total score for either objectively-scored or subjectively-scored criteria before the training (*Table 4-18*). This reinforces the observation that developing high-quality listening tasks was generally more challenging for the participants. Fewer tasks achieved a high total score on the subjectively-scored criteria (*Figure 4-12*) compared to the objectively-scored ones (*Figure 4-11*), which suggests that the subjectively-scored criteria were generally more difficult for the participants to meet.

Table 4-18. Descriptive statistics for the total scores on the objectively-scored and subjectively-scored criteria of B1 listening tasks

	Range	Min	Max	Mean		SD	Skewness		Kurtosis	
				Statistic	SE		Statistic	SE	Statistic	SE
Total scores on the objectively-scored criteria										
Pre-training	15	8	23	**18.48**	.71	3.55	-.90	.46	1.60	.90
Post-training	8	16	24	**20.68**	.41	2.08	-.57	.46	.17	.90
Total scores on the subjectively-scored criteria										
Pre-training	8	19	27	**23.56**	.44	2.22	-.43	.46	-.78	.90
Post-training	9	19	28	**24.40**	.46	2.33	-.70	.46	.10	.90

Post-training, the score range for the objectively-scored criteria was substantially lower than pre-training (from 15 to 8). While the score range for the subjectively-scored criteria is similar to the one pre-training, the distribution type changed: it was flat before the training, whereas it was peaked after the training, with more tasks gaining the total score of 24 or higher (20 tasks, compared to 14 pre-training). Moreover, while there were no tasks scoring the maximum possible on either set of criteria pre-training, there were three tasks that did so after it.

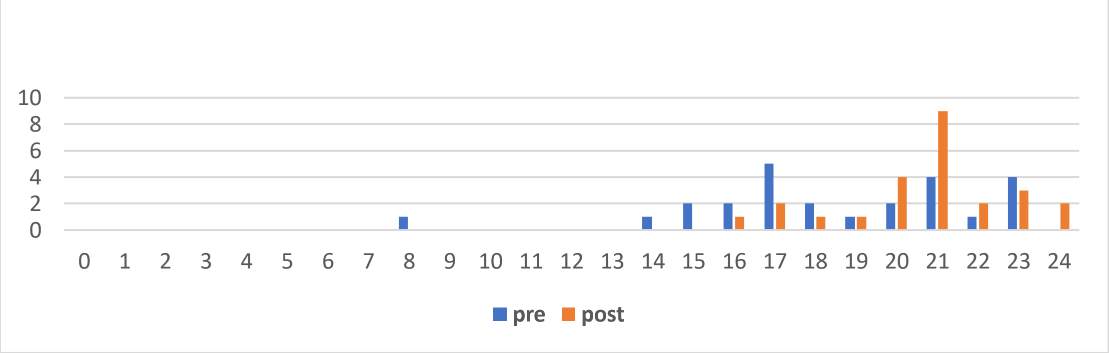


Figure 4-11. B1 listening tasks' objectively-scored criteria: descriptive statistics for the total scores

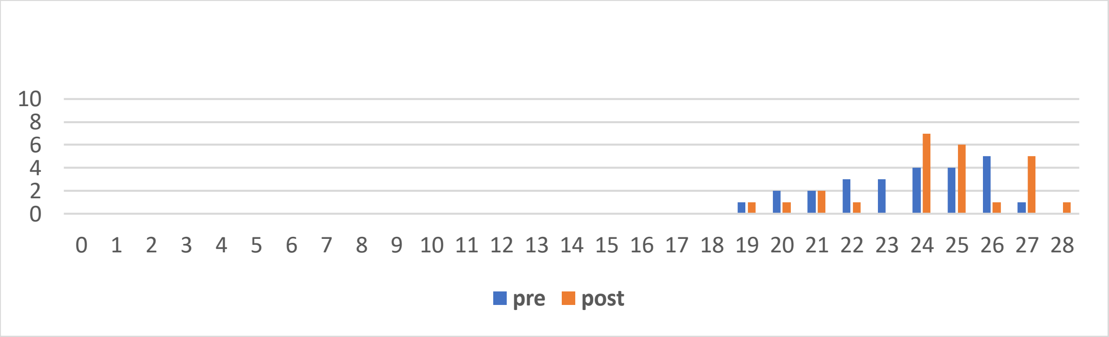


Figure 4-12. B1 listening tasks' subjectively-scored criteria: descriptive statistics for the total scores

Pre- and post-training scores on individual objectively-scored criteria

The 12 objectively-scored criteria used to evaluate B1 listening tasks are presented in *Table 4-19* (see also *Appendix 6).*

Table 4-19. Objectively-scored criteria to evaluate listening tasks

L1	Text: max. 300 words
L2	Text: lexis K1-K3 (1% of lexis can be proper names off frequency lists)
L3	Topic: From the list of topics for B1 level
L4	Function: From the list of functions for B1 level
L5	Items: 6 in total
L6	Items: either a set of notes or individual sentences
L7	Stem: Max 10 words including the key
L8	Stem: lexis K1-K2
L9	Stem: grammar A1-A2
L10	Stem: a paraphrase, i.e. does not literally repeat what is heard in the text
L11	Response: lexis K1-K2 (except for proper names that are spelt out, there should be no more than 1 item of this kind per task)
L12	Spelling / grammar / punctuation: correct, including the text, items and the key

A wider range of means for the objectively-scored criteria was found for the listening tasks – from 1.0 to 2.0 – compared to the other item types (*Table 4-20*). Participants were best able to conform to the text word-limit (L1, M=2.00) and item format (L5, M=2.00 and L6, M=1.92) requirements. At the same time, participants were struggling to meet the criteria on the choice of topic (L3, M=1.32), proofreading (L12, M=1.0), as well as two item-related criteria: grammatical complexity of the stem (L9, M=1.2) and the requirement for the stem to be a paraphrase of the input text (L10, M=1.28).

The mean values for most objectively-scored criteria were higher following the training, although there was still a wide range of mean scores, from 1.16 to 2.0. The largest increase in scores was observed for the vocabulary frequency requirements (L2 and L8), the requirement for stems to be paraphrases of the text (L10), and the proofreading requirement (L12). This observation is supported with the band frequency statistics: there were fewer band '0' scores (-7.4%) and band '1' scores (-3.6%), while there were more band '2' scores (+11%, see *Appendix 14*). However, there were two criteria that had lower mean values following the training: the input text word-limit requirement (L1) and the grammar requirement for item stems (L9).

Table 4-20. Descriptive statistics for the objectively-scored criteria of B1 listening tasks

Criteria	PRE-TRAINING					POST-TRAINING				
	Frequencies			Mean	SD	Frequencies			Mean	SD
	Band 0	Band 1	Band 2			Band 0	Band 1	Band 2		
L1	0	0	25	**2.00**	.00	0	4	21	**1.84**	.37
L2	2	9	14	**1.48**	.65	0	4	21	**1.84**	.37
L3	8	1	16	**1.32**	.94	1	4	20	**1.76**	.52
L4	5	2	18	**1.52**	.82	3	2	20	**1.68**	.69
L5	0	0	25	**2.00**	.00	0	0	25	**2.00**	.00
L6	1	0	24	**1.92**	.40	0	0	25	**2.00**	.00
L7	1	4	20	**1.76**	.52	2	1	22	**1.80**	.58
L8	2	11	12	**1.40**	.64	0	6	19	**1.76**	.44
L9	5	10	10	**1.20**	.76	6	9	10	**1.16**	.80
L10	5	8	12	**1.28**	.79	1	4	20	**1.76**	.52
L11	2	6	17	**1.60**	.64	2	6	17	**1.60**	.64
L12	9	7	9	**1.00**	.87	3	7	15	**1.48**	.71

Pre- and post-training scores on individual subjectively-scored criteria

The 14 subjectively-scored criteria used to evaluate B1 listening tasks are presented in *Table 4-21* (see also *Appendix 6).*

Table 4-21. Subjectively-scored criteria to evaluate B1 listening tasks

L13	Text: A monologue (recorded instructions, lectures, presentations, public announcements, TV/radio programmes, short talks, news reports).
L14	Text: sounds authentic according to the genre
L15	Text: accessible to a B1 level test-taker
L16	Text: the content is appropriate, culturally unbiased, not disturbing
L17	Text: suitable for testing, i.e. is NOT a parody, not silly, humorous, sarcastic, etc.
L18	Instruction: standard format is followed
L19	Items: test the ability to locate and record specific information from a monologue
L20	Items: do not test abilities unrelated to listening comprehension (e.g. maths, grammar, etc.)
L21	Items: each item (except for proper names that are spelt out) has one or two pieces of information in the text that act as a distractor
L22	Items: follow the order in the text
L23	Items: The necessary information for different items is distributed across the whole text with no two pieces of information appearing too close to each other in the text
L24	Stem: is clearly formulated in such a way that it restricts the number of possible correct answers
L25	Response: requires max. 3 words or a number heard in the text
L26	Response: All acceptable answers are included in the key

The subjectively-scored criteria (see *Table 4-21*) can be categorised into input text-related and item-related criteria. Pre-training, most participants coped well with three out of five text-related criteria (*Table 4-22*): suitability for testing (L17, M=1.92), accessibility at B1 proficiency level (L15, M=1.92), and fairness / lack of bias (L16, M=1.8). The text genre criterion (L13) had a lower mean value (1.68), and the text authenticity criterion (L14) had by far the lowest mean value (0.84) of all subjectively-scored criteria. The item-related criteria L19 and L20, which are concerned with the task construct, achieved generally high scores (M=1.88 and 1.92 respectively). The latter is an indication that most participants were able to operationalise the intended construct in items. Of the four lowest-scoring criteria three were item-related: the requirement for each item to have distracting information in the text (L21, M=1.28), for the stem to be clearly formulated (L24, M=1.28), and for the key to include all acceptable answers (L26, M=1.44).

Table 4-22. Descriptive statistics for the subjectively-scored criteria of B1 listening tasks

Criteria	PRE-TRAINING					POST-TRAINING				
	Frequencies			Mean	SD	Frequencies			Mean	SD
	Band 0	Band 1	Band 2			Band 0	Band 1	Band 2		
L13	2	4	19	**1.68**	.63	1	3	21	**1.80**	.50
L14	6	17	2	**.84**	.55	0	15	10	**1.40**	.50
L15	0	2	23	**1.92**	.28	0	3	22	**1.88**	.33
L16	0	5	20	**1.80**	.41	0	4	21	**1.84**	.37
L17	1	0	24	**1.92**	.40	0	0	25	**2.00**	.00
L18	0	4	21	**1.84**	.37	0	4	21	**1.84**	.37
L19	1	1	23	**1.88**	.44	2	3	20	**1.72**	.61
L20	0	2	23	**1.92**	.28	0	2	23	**1.92**	.28
L21	6	6	13	**1.28**	.84	8	4	13	**1.20**	.91
L22	0	0	25	**2.00**	.00	0	0	25	**2.00**	.00
L23	1	2	22	**1.84**	.47	0	1	24	**1.96**	.20
L24	3	12	10	**1.28**	.68	0	16	9	**1.36**	.49
L25	0	2	23	**1.92**	.28	0	0	25	**2.00**	.00
L26	2	10	13	**1.44**	.65	0	13	12	**1.48**	.51

Following the training, eight subjectively-scored criteria had higher mean values than pre-training. The biggest improvement was seen in the participants' ability to produce texts of the required genre (L13), distribute targeted information evenly throughout the text (L23), write clear items (L24) and, especially, write authentic-sounding input texts (L14). The mean score for the latter criterion was 0.84 pre-training but 1.4 post-training: no task scored '0' on this criterion post-training (compared to six pre-training), and ten tasks scored '2' (compared to two pre-training). Although the mean values for text suitability (L17) and item response characteristics (L25) were only slightly higher following the training, from 1.92 to 2.0, the difference is nevertheless meaningful because it testifies that all item writers mastered these two criteria following the training.

At the same time, three criteria had lower mean values after the training, compared to before. They are text accessibility (L15), construct (L19), and distractor (L21) requirements. The distractor requirement scored second lowest pre-training (M=1.28) and was the absolute lowest post-training (M=1.20). This is similar to what was observed for the C1 grammar items: the mean value on the requirement that distractors are strong and plausible was lower following the training and in fact the lowest of all mean values for the subjectively-scored criteria. It seems that novice item writers in this study faced a continuous struggle to produce distractors for different item types. It should be noted, though, that the A2 grammar items were awarded substantially higher scores on the same requirement after training, which might indicate that this requirement's difficulty is linked to the proficiency level of the items.

In terms of band frequency statistics (see *Appendix 14*), there were somewhat fewer band '0' scores (-3.1%) and somewhat more band '2' scores (+2.8%) following the training. However, unlike for other item types, the listening tasks saw only marginally more band '1' scores for the subjectively-scored criteria post-training (from 19.1% to 19.4%).

4.2.2 Findings from the Wilcoxon signed-rank test analyses

Tables 4-23 to 4-31 present the results of the Wilcoxon signed-rank tests performed on the raw item evaluation scores: Z-scores, asymptotic significance values (*p*), and effect sizes (*r*). The significance level set at p<.05 and interpreted using Cohen's (1988) guidelines: values ≤ 0.3 are viewed as a small effect size, 0.3 to 0.5 represent a medium effect size, and ≥0.5 a large effect size. Statistically significant results are colour-coded in the tables: green indicates that scores for the items produced after the training were significantly higher than scores for the items produced before the training. When the opposite was true, that is post-training scores were significantly lower than the pre-training ones, the results are highlighted in red.

Findings on the A2 and C1 grammar items

Table 4-23 shows that the total scores for both the A2 and C1 grammar items were statistically significantly higher following the training (*p*=0.01, *r*=0.38 for A2 items; *p*=0.01, *r*=0.39 for C1 items). This was mainly because the total scores for the objectively-scored criteria were significantly higher following the training for both the A2 (*p*=0.00, *r*=0.46) and C1 (*p*=0.00, *r*=0.42) items. At the same time, a comparison of the pre- and post-training total scores for the subjectively-scored criteria, as well as the scores on the overall acceptability criterion, did not show statistically significant differences.

Table 4-23. Wilcoxon signed-rank test results for the score totals and the overall acceptability criterion of A2 and C1 grammar items

	Overall total	Objectively-scored criteria total	Subjectively-scored criteria total	Overall acceptability criterion
A2 grammar				
Z-score	-2.71	-3.26	-1.09	-1.29
Asymp. Sig. (2-tailed)	0.01*	0.00*	0.27	0.19
Effect size	0.38	0.46	0.15	0.18
C1 grammar				
Z-score	-2.78	-2.94	-0.77	0
Asymp. Sig. (2-tailed)	0.01*	0.00*	0.44	1.00
Effect size	0.39	0.42	0.11	0

Scores on three objectively-scored criteria, both for the A2 and C1 items, were significantly higher following the training (*Table 4-24*). These comprise the requirement to integrate repeating words into the stem (G5, *p*=0.02, *r* =0.33), for the key to be indicated (G6, *p*=0.01, *r*=0.42), and for the item to be proofread (G10, *p*=0.02, *r*=0.33), all of medium effect size.

Table 4-24. Wilcoxon signed-rank test results for objectively-scored criteria of A2 and C1 grammar items

	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
A2 grammar										
Z-score	-1	0	0	0	-2.31	-2.44	-0.33	-1.51	-1.82	-2.33
Asymp. Sig. (2-tailed)	0.32	1.00	1.00	1.00	0.02*	0.01*	0.73	0.13	0.06	0.02*
Effect size	0.14	0	0	0	0.33	0.42	0.05	0.21	0.26	0.33
C1 grammar										
Z-score	0	-0.81	0	-1	-2.61	-2.12	-0.57	-1.61	-0.75	-2.73
Asymp. Sig. (2-tailed)	1.00	0.41	1.00	0.31	0.01*	0.03*	0.56	0.10	0.45	0.01*
Effect size	0	0.11	0	0.14	0.37	0.30	0.08	0.23	0.11	0.39

As for the subjectively-scored criteria (*Table 4-25*), the post-training scores were significantly higher on the requirement for distractors to be grammatically correct as a stand-alone (G14) for the A2 (*p*=0.01, *r*=0.38) and C1 (*p*=0.03, *r*=0.30) grammar items, with medium effect sizes.

At the same time, only A2 item scores were significantly higher post-training on the requirement for distractors to be strong and plausible (G12, *p*=0.04, *r*=0.29), with a small-to-medium effect size.

Table 4-25. Wilcoxon signed-rank test results for subjectively-scored criteria of A2 and C1 grammar items

	G11	G12	G13	G14	G15	G16	G17	G18
A2 grammar								
Z-score	-0.67	-2.02	-1.41	-2.69	-0.03	-0.27	0	0
Asymp. Sig. (2-tailed)	0.49	0.04*	0.15	0.01*	0.97	0.78	1.00	1.00
Effect size	0.09	0.29	0.19	0.38	0	0.04	0	0
C1 grammar								
Z-score	-1.42	-0.72	-1	-2.12	-1.15	-1.66	-1	0
Asymp. Sig. (2-tailed)	0.15	0.46	0.31	0.03*	0.24	0.09	0.31	1.00
Effect size	0.20	0.10	0.14	0.30	0.16	0.23	0.14	0

Findings on the B2 writing prompts

A comparison of pre- and post-training total scores, as well as scores on individual criteria for B2 writing prompts produced no statistically significant results (*Table 4-26, Table 4-27, Table 4-28*).

Table 4-26. Wilcoxon signed-rank test results for score totals and overall acceptability of B2 writing prompts

	Objectively-scored criteria total	Subjectively-scored criteria total	Overall total	Overall acceptability criterion
Z-score	-1.21	-0.66	-.89	-0.30
Asymp. Sig. (2-tailed)	0.22	0.50	.37	0.76
Effect Size	0.17	0.09	0.12	0.04

Table 4-27. Wilcoxon signed-rank test results for objectively-scored criteria of B2 writing prompts

	W1	W2	W3	W4	W5
Z-score	-0.27	0	-0.81	-1.26	-1.23
Asymp. Sig. (2-tailed)	0.78	1.00	0.41	0.20	0.21
Effect size	0.04	0	0.11	0.18	0.17

Table 4-28. Wilcoxon signed-rank test results for subjectively-scored criteria of B2 writing prompts

	W6	W7	W8	W9	W10	W11	W12	W13	W14	W15	W16
Z-score	-0.90	-0.81	-0.44	-0.53	-0.81	0	-1.63	-1.34	-0.36	-0.44	-1.13
Asymp. Sig. (2-tailed)	0.36	0.41	0.65	0.59	0.41	1.00	0.10	0.18	0.71	0.65	0.25
Effect size	0.13	0.11	0.06	0.07	0.11	0	0.23	0.19	0.05	0.06	0.16

Findings on the B1 listening tasks

The total scores for B1 listening items (*Table 4-29*) were significantly higher after the training (*p*=0.00, *r*=0.43). Similar to the grammar items, this was largely due to the objectively-scored criteria – their score totals were significantly higher after the training, with a medium effect size (*p*=0.00, *r*=0.42), while a comparison of score totals on the subjectively-scored criteria did not produce significant results.

Table 4-29. Wilcoxon signed-rank test results for score totals and overall acceptability of B1 listening tasks

	Overall total	Objectively-scored criteria total	Subjectively-scored criteria total	Overall acceptability criterion
Z-score	-3.08	-2.95	-1.35	-0.63
Asymp. Sig. (2-tailed)	0.00*	0.00*	0.18	0.53
Effect size	0.43	0.42	0.19	0.09

Scores on four objectively-scored criteria were significantly higher after the training (*Table 4-30*), with medium effect size: two vocabulary frequency-related criteria (L2 and L8, *p*=0.01, *r*=0.35), the requirement for items to be a paraphrase of information in the input text (L10, *p*=0.03, *r*=0.31), and the proofreading requirement (L12, *p*=0.02, *r*=0.32). At the same time, scores on the text word-limit criterion were significantly lower after the training compared to before (L1, *p*=0.046, *r*=0.28), although the effect size was small.

Table 4-30. Wilcoxon signed-rank test results for objectively-scored criteria of B1 listening tasks

	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12
Z-score	-2	-2.49	-1.95	-0.73	0	-1	-0.14	-2.49	-0.22	-2.19	0	-2.24
Asymp. Sig. (2-tailed)	0.046*	0.01*	0.05	0.46	1.00	0.31	0.88	0.01*	0.82	0.03*	1.00	0.02*
Effect size	0.28	0.35	0.27	0.10	0	0.14	0.02	0.35	0.03	0.31	0	0.32

As for the subjectively-scored criteria (*Table 4-31*), scores on only one of them - the text authenticity criterion - were significantly higher after the training, with a medium effect size (*p*=0.00, *r*=0.43).

Table 4-31. Wilcoxon signed-rank test results for subjectively-scored criteria of B1 listening tasks

	L13	**L14**	**L15**	**L16**	**L17**	**L18**	**L19**	**L20**	**L21**	**L22**	**L23**	**L24**	**L25**	**L26**
Z-score	-0.72	-3.07	-0.44	-0.33	-1	0	-1.41	0	-0.44	0	-1.13	-0.53	-1.41	-0.24
Asymp. Sig. (2-tailed)	0.47	0.00*	0.65	0.73	0.31	1.00	0.15	1.00	0.65	1.00	0.25	0.59	0.15	0.80
Effect size	0.10	0.43	0.06	0.05	0.14	0	0.19	0	0.06	0	0.16	0.07	0.19	0.03

4.2.3 Findings from the gain ratio statistics

As discussed in the Methodology Chapter (Section 3.5.1.4), Wilcoxon signed-rank tests have limitations in their application to this study's data. This is because some items received high scores on a number of evaluation criteria prior to the training already, which left a limited scope for change following the training. A small change in a total score might pass undetected by a statistical test but, in operational testing, it might make all the difference between an item being returned for revision or accepted for live testing. Therefore, in addition to the Wilcoxon signed-rank tests, the gain ratio technique was used to provide insights into more nuanced changes in the quality of items from before to after the training. Furthermore, the gain ratio technique allowed insights into item-writer variation within the training cohort, as the statistics were obtained individually for each participant. Gain ratio statistics for each individual participant are reported in *Tables 4-32 to 4-35.*

Table 4-32. Gain ratio statistics for A2 grammar items

	Sum of scores on objectively-scored criteria			Sum of scores on subjectively-scored criteria		
	pre	**post**	**GR**	**pre**	**post**	**GR**
Josh	20	20	N/A[6]	12	15	75%
Henry	13	19	86%	13	13	0%
James	12	19	87%	13	15	67%
Ted	18	18	0%	11	14	60%
Alex	20	18	loss[7]	14	15	50%
Joe	19	20	100%	13	12	loss
Daniel	14	18	67%	14	15	50%
Arthur	18	20	100%	11	16	100%
Lucas	18	19	50%	15	14	loss
Emily	19	20	100%	14	10	loss
Logan	18	20	100%	9	14	71%
Adam	17	20	100%	8	16	100%
Olivia	19	20	100%	9	14	71%
Chloe	14	20	100%	14	15	50%
Lucy	17	20	100%	14	14	0%
Jake	14	14	0%	15	10	loss
Mathew	20	20	N/A	14	13	loss
Liz	19	19	0%	16	15	loss
Rose	18	20	100%	14	11	loss
Luke	17	18	33%	11	16	100%
Stanley	16	20	100%	16	15	loss
Austin	19	19	0%	14	14	0%
Nathan	20	20	N/A	14	14	0%
Mason	19	19	0%	16	15	loss
Ryan	18	20	100%	15	14	loss

[6] The gain ratio statistic cannot be calculated due to both pre- and post-training items gaining the maximum total score, i.e. no gain is possible.

[7] The gain ratio statistic cannot be calculated because the post-training total score is smaller than the pre-training one, i.e. there is no gain but a loss in item quality following the training.

Table 4-33. Gain ratio statistics for C1 grammar items

	Sum of scores on objectively-scored criteria			**Sum of scores on subjectively-scored criteria**		
	pre	**post**	**GR**	**pre**	**post**	**GR**
Josh	20	19	loss	11	14	60%
Henry	15	20	100%	12	15	75%
James	12	20	100%	7	15	89%
Ted	15	18	60%	15	14	loss
Alex	20	16	loss	13	14	33%
Joe	20	19	loss	12	16	100%
Daniel	18	20	100%	16	13	loss
Arthur	20	20	N/A	16	12	loss
Lucas	14	20	100%	12	14	50%
Emily	17	19	67%	12	13	25%
Logan	18	17	loss	11	13	40%
Adam	16	20	100%	14	15	50%
Olivia	19	16	loss	15	14	loss
Chloe	12	19	87%	15	14	loss
Lucy	16	20	100%	13	14	33%
Jake	14	17	50%	16	16	N/A
Mathew	15	20	100%	15	14	loss
Liz	19	19	0%	14	12	loss
Rose	20	20	N/A	16	15	loss
Luke	16	19	75%	15	13	loss
Stanley	14	20	100%	12	13	25%
Austin	17	18	33%	13	14	33%
Nathan	17	20	100%	16	15	loss
Mason	20	19	loss	14	16	100%
Ryan	15	20	100%	15	14	loss

Table 4-34. Gain ratio statistics for B2 writing prompts

	Sum of scores on objectively-scored criteria			**Sum of scores on subjectively-scored criteria**		
	pre	**post**	**GR**	**pre**	**post**	**GR**
Josh	6	10	100%	16	20	67%
Henry	7	8	33%	18	21	75%
James	8	10	100%	21	21	0%
Ted	8	10	100%	9	22	100%
Alex	8	7	loss	21	19	loss
Joe	10	10	N/A	21	18	loss
Daniel	9	9	0%	21	21	0%
Arthur	9	10	100%	21	20	loss
Lucas	10	9	loss	19	18	loss
Emily	10	9	loss	22	22	N/A
Logan	9	10	100%	21	22	100%
Adam	9	10	100%	19	21	67%
Olivia	10	8	loss	22	22	N/A
Chloe	9	9	0%	22	22	N/A
Lucy	8	10	100%	21	21	0%
Jake	6	8	50%	21	21	0%
Mathew	10	8	loss	20	22	100%
Liz	10	10	N/A	22	20	loss
Rose	7	10	100%	21	19	loss
Luke	8	6	loss	21	20	loss
Stanley	10	10	N/A	19	19	0%
Austin	10	10	N/A	19	21	67%
Nathan	10	9	loss	20	19	loss
Mason	10	9	loss	21	22	100%
Ryan	8	10	100%	20	22	100%

Table 4-35. Gain ratio statistics for B1 listening tasks

	Sum of scores on objectively-scored criteria			**Sum of scores on subjectively-scored criteria**		
	pre	**post**	**GR**	**pre**	**post**	**GR**
Josh	15	19	44%	20	24	50%
Henry	23	21	loss	21	27	86%
James	17	21	57%	26	27	50%
Ted	16	22	75%	22	24	33%
Alex	15	17	22%	19	26	78%
Joe	23	24	100%	23	25	40%
Daniel	8	17	56%	26	27	50%
Arthur	21	21	0%	20	25	62%
Lucas	17	20	43%	25	19	loss
Emily	17	22	71%	24	22	loss
Logan	18	23	83%	26	25	loss
Adam	19	20	20%	21	24	43%
Olivia	22	22	0%	24	25	25%
Chloe	17	21	57%	23	21	loss
Lucy	14	21	70%	22	20	loss
Jake	16	21	62%	24	27	75%
Mathew	17	22	71%	26	28	100%
Liz	20	20	0%	25	21	loss
Rose	23	21	loss	22	24	33%
Luke	18	16	loss	25	25	0%
Stanley	21	20	loss	23	24	20%
Austin	21	20	loss	25	24	loss
Nathan	23	23	0%	27	27	0%
Mason	21	24	100%	26	24	loss
Ryan	20	23	75%	24	25	25%

The gain ratio statistics supported the findings from the Wilcoxon signed-rank tests that, after the training, many participants produced higher-quality items on the objectively-scored criteria. There were 16 participants for the A2 grammar items (*Table 4-32*), 16 participants for the C1 grammar items (*Table 4-33*), 11 participants for the writing prompts (*Table 4-34*), and 16 participants for the listening tasks *(Table 4-35)* whose post-training items demonstrated gains on the sum of the objectively-scored criteria. For the majority of them, the gain was over 50%, with 11 (A2 grammar items), 10 (C1 grammar items), and nine (B2 writing prompts) participants achieving 100% gain, which means that the items were awarded the maximum possible score on the sum of the objectively-scored criteria following the training. For the listening tasks, the instances of 100% gain were fewer with only two participants achieving it. Additionally, some participants gained the maximum total scores on the sum of the objectively-scored criteria for both their pre- and post-training items: three participants for the A2 grammar items, two participants for the C1 grammar items, and four participants for the writing prompts. For the listening tasks, no participant scored the maximum on the sum of the objectively-scored criteria both before and after the training.

There were also several instances of loss in item quality on the objectively-scored criteria, which means that the post-training item was awarded a lower score on the sum of the objectively-scored criteria compared to the corresponding pre-training one: there was one instance for the A2 grammar items (*Table 4-32*), six for the C1 grammar items *(Table 4-33*), eight for the writing prompts (*Table 4-34*), and five for the listening tasks (*Table 4-35*). Notably, the loss was observed mostly for those items that had already achieved the maximum or near-maximum total score before the training. Moreover, the loss was normally by one point only; for example, Alex scored '20' out of 20 on the objectively-scored criteria for his C1 grammar item before the training, but '19' after the training. For the writing prompts, the instances of loss were predominantly observed because of a lower score gained on the input message word-limit criterion (W1), which also had a slightly lower post-training mean score; for the listening tasks, the instances of loss mostly happened due to a lower score gained on the text word-limit criterion (L1), which post-training mean score was significantly lower compared to the pre-training one.

Compared to what was found for the objectively-scored criteria, somewhat fewer item writers produced better-quality items on the subjectively-scored criteria following the training: 11 for the A2 items (*Table 4-32*), 13 for the C1 items (*Table 4-33*), nine for the writing prompts (*Table 4-34*), and 15 for the listening tasks (*Table 4-35*). Another difference was that, while there were many instances of 100% gain on the objectively-scored criteria, there were much fewer instances of 100% gain on the subjectively-scored criteria: three for the A2 grammar items, two for the C1 grammar items, five for the writing prompts, and one for the listening tasks.

Moreover, compared to the findings on the objectively-scored criteria, there were considerably more instances of loss in item quality on the subjectively-scored criteria following the training: 10 for the A2 items (*Table 4-32*), 11 for the C1 items (*Table 4-33*), eight for the writing prompts (*Table 4-34*), and eight for the listening tasks (*Table 4-35*). Many participants whose pre-training grammar items gained very high total scores on the sum of the subjectively-scored criteria, produced lower-scoring items on the subjectively-scored criteria following the training. For four of them, the loss occurred on both A2 and C1 items, while for the rest the loss was observed on one item only. For the writing prompts, 19 out of 25 item writers scored high on the sum of the subjectively-scored criteria before the training, eight of those participants produced lower-quality items on the subjectively-scored criteria following the training. The same pattern, whereby a participant scored high on the sum of the subjectively-scored criteria before the training but then scored lower after the training, was also observed for eight participants with regard to their listening tasks. In most instances, the sum of scores decreased by a small margin only.

Overall, four participant profiles (*Table 4-36*) emerged through a comparison of their pre- and post-training item evaluations on the subjectively-scored criteria: (a) a small number of participants whose pre-training items scored the lowest on the sum of the subjectively-scored criteria (so-called *'outliers'*) but who produced much higher quality items after the training; (b) participants who produced good quality items before the training, and whose post-training items scored even higher (so-called *'high-achievers'*); also included in this group are those participants whose items scored the maximum or near-maximum on both occasions; (c) participants whose pre-training items scored high (80% or more of the maximum score on the sum of the subjectively-scored criteria) but whose post-training items scored slightly lower; (d) all other participants who displayed more unique trends that could not be categorised. For example, their item quality improved following the training, but the improvement was not as drastic as for *'outliers'* or the post-training scores were not as high as for *'high-achievers'*. Alternatively, the loss in quality on their post-training items was larger than for profile C participants. *Table 4-36* provides the number of participants for each category. The numbers vary depending on item type: for instance, there were more *'high-achievers'* with regard to writing and listening items, while there were more participants whose post-training grammar items scored lower on the sum of the subjectively-scored criteria, compared to the pre-training ones.

Table 4-36. Four trainee profiles

	Profile A	***Profile B***	***Profile C***	***Profile D***
	'Outliers'	'High-achievers'	Lower scores post-training	Others
Grammar A2	3	7	10	5
Grammar C1	1	8	11	5
Writing B2	2	14	8	1
Listening B1	3	11	8	3

Finally, a comparison of item-writer performance on the objectively-scored and subjectively-scored criteria demonstrated few correlations - there were a number of cases with a gain on the objectively-scored criteria but a loss or zero gain on the subjectively-scored criteria, or vice versa, for the same item. A comparison across item types revealed a similar situation; for example, a gain in scores for the grammar items did not guarantee a gain in scores for the listening task. Only one participant, Adam, demonstrated gains for all item types on both the objectively-scored and the subjectively-scored criteria. Several participants had gains for most item types on both sets of criteria, among them Josh, Henry, James, Ted, Daniel, Logan, and Ryan. Only one participant, Liz, produced items that generally scored lower post-training compared to her pre-training items. However, no trend could be identified for the remaining participants. For example, Lucas showed gains for the objectively-scored criteria of most items but losses for the relevant subjectively-scored criteria. For Alex, the opposite was true. Some other participants demonstrated an even greater mix of results.

4.2.4 Summary of the quantitative findings

Below, I integrate and summarise the quantitative findings from the descriptive statistics, Wilcoxon signed-rank tests, and gain ratios.

Summary of the findings for the pre-training item evaluations

Prior to the training, the participants already had some ability to conform to the item specification requirements, which is supported by the score frequency analysis: band ‘2’ score counts were higher than band ‘1’ and band ‘0’ counts for all item types. Also, more band ‘1’ than band ‘0’ scores were awarded, which means that, in principle, the majority of lower-quality items could be improved through revision rather than having to be rejected out-of-hand. At the same time, there was a small number of items for each item type which scored much lower than the rest, that is several participants demonstrated a much lower item-writing ability compared to their peers for that item type. Notably, these participants were different for each item type, except for Josh whose writing prompt and listening task were both outliers. Many pre-training items scored band ‘2’ on a range of individual criteria; however, few items achieved band ‘2’ on all criteria, with no listening task doing so. In other words, although many of the items produced pre-training had merit, most would not be accepted for live testing without revision. This observation is supported by the scores on the overall item acceptability criterion: the mean values on this criterion were much lower than the mean values on most individual criteria for each item type.

Pre-training, the participants were generally better at developing writing prompts than grammar items or listening tasks – no mean value for any writing prompt criterion was below 1.6, while some much lower mean values were observed for grammar and especially listening items. The acceptability rate for listening items was also lower than for the other three item types.

More participants were successful at meeting each individual criterion than those who failed. The two notable exceptions were the text authenticity criterion for the listening tasks (L14) and the distractor criterion for the A2 grammar items (G12), with over 50% of participants having failed to meet these criteria requirements pre-training. The requirement for both grammar and listening items to have strong and plausible distractors was particularly difficult for the untrained item writers. Participants also had problems with the clarity and conciseness of items (grammar and listening), the clarity of instructions (writing), and, among the objectively-scored criteria, the requirement to proofread items before submission (all item types). At the same time, most untrained participants were able to conform to the criteria that concerned word-limit, vocabulary frequency, and item format among the objectively-scored criteria; fairness, lack of bias, and the suitability of content for testing among the subjectively-scored criteria.

Summary of the findings for the post-training item evaluations

Fourteen and 19 participants produced better-quality A2 and C1 grammar items, respectively, and 18 participants produced better quality listening tasks following the training, as evident from the Wilcoxon signed-rank test results: overall post-training score for the A2 and C1 grammar items and for the B1 listening tasks were statistically significantly higher. The improvement in the overall item quality, however, was in a large part due to the improvement in quality on the objectively-scored criteria, for which the post-test total scores were statistically significantly higher than the pre-test ones at p=.00 level. At the same time, the changes in the total scores on the subjectively-scored criteria were not statistically significant. For the B2 writing prompts, Wilcoxon signed-rank tests detected no significant difference in the overall scores or the total scores for the objectively/subjectively-scored criteria following the training. However, most B2 writing prompts already scored quite high prior to the training, so differences in scores might not have been easy to detect.

There were no outliers (i.e. the items that scored much lower than the rest) among the post-training items, except for one A2 grammar item, while there were several outliers pre-training for each item type. This suggests that participants' item-writing ability was more uniform across the whole cohort, following the training. However, the mean scores on the overall acceptability criterion were still low (below 1.50 for all item types), with the number of items that could be immediately accepted for live testing being somewhat higher for the A2 grammar items (11), lower for the C1 grammar items (7) and the writing prompts (8), and the lowest for the listening tasks (3). It seems that, following the training, most items still needed further revision, with the listening tasks posing the greatest difficulty to the item writers.

Analysis of the descriptive statistics revealed that the mean scores for most objectively-scored criteria of all item types were higher after the training, with many nearing or equalling the maximum possible score. However, Wilcoxon signed-rank tests revealed a significant difference only in three objectively-scored criteria for the grammar items (the same ones for the A2 and C1 items) and four objectively-scored criteria for the listening tasks. For all other criteria the post- vs pre-training score difference might not have been large enough to be detected because the corresponding pre-training scores were already quite high. However, there was also one objectively-scored writing prompt criterion (the input message word-limit) and two listening task criteria (the text word-limit and the grammar of item stems) that had a decrease in the mean score following the training. For the listening text word-limit criterion (L1) the decrease was so substantial that the Wilcoxon signed-rank test showed it as statistically significant.

Compared to the objectively-scored criteria, Wilcoxon signed-rank tests identified fewer subjectively-scored criteria with statistically significantly higher scores following the training: two distractor-related criteria for the A2 grammar items, one distractor-related criterion for the C1 grammar item, and one criterion, the input text's authenticity, for the listening tasks. There was no significant difference in the scores on any individual criterion for the writing prompts. The statistically significant results for the distractor-related criteria of the grammar items suggest that the participants' ability to produce distractors – the area of biggest concern pre-training – improved following the training. However, the requirement to create strong and plausible distractors still posed the greatest difficulty even post-training: the relevant C1 mean score was somewhat lower compared to the pre-training one and, even though the A2 mean score was statistically significantly higher compared to the pre-training one, it was still the lowest among the mean scores for all subjectively-scored criteria. Distractors also posed great difficulty to participants with regard to the listening tasks: one distractor-related criterion had a lower mean score following the training. Another area of difficulty was construct-related requirements: one construct-related criterion saw a decrease in scores for the listening tasks and two - for the A2 grammar items (but not the C1 items). Furthermore, it seems that A2 and C1 grammar items posed different challenges to the participants: construct-related scores decreased for the A2 items while distractor-related scores decreased for the C1 items.

The gain ratio statistics supported the observation that, overall, there was a more uniform improvement in the quality of items on the objectively-scored criteria across the participants cohort. There were more instances of gain and fewer instances of loss on the objectively-scored criteria, compared to the subjectively-scored ones, which is true for all item types. With regard to the pre- vs post-training item quality of items on the subjectively-scored criteria, four participant profiles emerged: profile A participants produced lowest-scoring items on the sum of the subjectively-scored criteria pre-training but they wrote much higher quality items following the training; profile B participants produced high quality pre-training items while their post-training items were of similarly high or higher quality; profile C participants' pre-training items scored high, but their post-training items scored slightly lower; profile D includes all the remaining participants whose item quality improved following the training but who did not fall into any of the three categories above.

The analysis of gain ratio statistics revealed no correlations across item types: for example, one and the same participant could be in profile A for the grammar items, in profile B for the writing prompt, and in profile C for the listening task. Moreover, conflicting results were often observed for A2 and C1 items: those item writers whose A2 item quality improved following the training did not necessarily perform equally well on their C1 items, and vice versa.

4.3 Item-writing skill development (RQ2)

This section reports on findings related to the second research question: "*How did the participants' item-writing skills develop through the training, as perceived by the participants in interviews?"* It draws on the qualitative data from the interviews that were conducted with the participants on completion of their pre- and post-training item-writing assignments. The findings are reported in two sections: findings relevant to all item types focused on in the assignments are presented in Section 4.3.1, while item type-specific findings are reported in Section 4.3.2 which consists of three sub-sections: grammar items (4.3.2.1), writing prompts (4.3.2.2), and listening tasks (4.3.2.3). The findings are summarised in Section 4.3.3.

4.3.1 Findings relevant to all item types

In the interviews, most participants dwelt on item-writing difficulty (4.3.1.1), their attitude to item-writing (4.3.1.2), and their use of the test specifications, including example items, during the item-writing process (4.3.1.3). Although many comments were item-specific, many were also made about item-writing as a whole. Moreover, what participants said about one item type was often repeated for a different item type, for instance with regard to their use of the specifications. Therefore, I discuss these three topics in one section rather than present them separately for each item type.

4.3.1.1 Perceived difficulty of item-writing

Before the training

All 17 participants who were interviewed pre-training talked about how difficult it was for them to write the items, saying it was *'difficult', 'not easy', 'hard', challenging',* or *'tricky.'* There were many reasons why participants found item-writing difficult. For example, it was:

- harder than they had expected (Arthur)
- difficult to find time to write (Adam)
- hard working alone, as opposed to working in groups (Josh)
- difficult '*com*[ing] *up with lots of ideas'* (Ted, Nathan)
- difficult to understand the CEFR proficiency levels (Olivia)

For specific item types, participants found the listening task by far the most difficult. Seven participants commented that the task was generally hard to write, while others mentioned specific difficulties such as creating an authentic-sounding text (Arthur), thinking of an appropriate situation to write about (Ted), and keeping to the specified word count and vocabulary frequencies (Ted). Developing gap-fill items and creating distractors was also difficult for some. Moreover, Josh complained that the listening task specifications were very long and difficult to digest.

Grammar items came second according to the number of comments on difficulty. Some participants complained that the *'specifics'* of grammar were difficult to understand for them as native speakers. Other difficulties mentioned included observing the word count, creating stems, and choosing the right topic.

The writing prompt was reported to be the least challenging. Daniel mentioned it was hard for him to find *'the right scenario'*, while some other participants mentioned it was challenging to keep to the syntactic (Nathan) and lexical (James and Olivia) specifications. Jake said it was not particularly difficult for him to create one writing prompt, but he could envisage it would be hard to come up with many prompts for the same set of specifications.

After the training

Post-training, ten participants said it was easier for them to produce items. However, it was only *'slightly'* or '*a bit'* easier, with none of the participants thinking that item-writing was very easy or straightforward, even after the training. Quotes from some participants can help explain why:

> *The first time it took me a while because I didn't know what I was doing. And the second time it took me a while because I did know what I was doing and I had to check the specs all the time* (Adam)
>
> *It was easier in one respect because I have more experience with item writing, but it was more difficult because in fact when I did it the first time … I wasn't being as careful as I am now* (Arthur)
>
> *I don't feel that writing these tasks is any easier ... It's just that the approach is a bit more clear* (Joe)
>
> *I found it easier but it's not easy … I knew what I was doing a bit more this time* (James)

Generally, most participants thought that 'easy' was not the right word to use for item-writing, which is a very labour-intensive activity requiring a lot of attention to detail. However, many said they felt more '*confident*' or '*comfortable*' doing item-writing after they completed the training (see Section 4.4.2 for a discussion of the training's role in item-writing skill development):

> *I had a sense of confidence this time that I didn't have the first time because I had an idea of what I needed to do* (Daniel)
>
> *I feel fairly comfortable. It's difficult, still difficult, but I think that will ease over time* (Henry)

Like Henry, other participants also felt that the training was only the first step and they needed considerably more time to fully develop into professional item writers. For example, Emily said that *"you need to write quite a lot of items before you can say you actually find it easier."*

In terms of specific item types, 17 participants (out of 19 interviewed), found the listening task most difficult to write. The listening task was perceived as most time-consuming and complex, with "*an awful lot to keep in your head while you're writing it*" (Ted). One difference with pre-training, however, was that more participants found producing the text more difficult than writing the gap-fill items, with Mason saying that after the training *"the items were easier to write and the text was harder".*

Four participants felt that producing the grammar items was substantially easier for them after the training. Twelve other participants, though, still found grammar items challenging to write. Three difficulties were named: producing lower-proficiency level items, targeting the right construct, and thinking of strong and plausible distractors. Notably, these challenges were different from the ones named before the training, when participants mostly raised issues with meeting the objectively-scored criteria (more on this in Section 4.3.2).

Similar to pre-training, most participants found the writing prompts the easiest to produce. All six participants who mentioned a difficulty stated that finding '*the right scenario'* was the biggest challenge. Only one participant, Jake, said that the writing prompt was the hardest to produce because, in his opinion, "*it just seems a rather unnatural and unrealistic task"*.

4.3.1.2 Attitudes to item-writing

Before the training

Despite the many difficulties that participants had with writing the items, positive attitudes prevailed. Pre-training, participants expressed their positive attitude to the item-writing activity 50 times in the interviews, saying it was '*interesting', 'fun', 'enjoyable', 'nice', 'exciting',* as well as *'challenging'* used with a positive connotation. This enthusiasm towards item-writing varied among participants, with Logan mentioning it 13 times, Daniel eight times, Lucy and Olivia seven times each, and others only once or twice during the interview.

Most participants liked item-writing in general, without referring to a particular item type. What they enjoyed most was the process of item-writing (James), the fact they had to be creative (Lucy), and that it made their brain work hard (Daniel, Lucy). Mason said he *"enjoyed pushing against the restraints of the challenges that you get".* Surprisingly, the listening task, which had been identified as the most difficult, was at the same time thought to be the most interesting to write. Some participants said it was interesting exactly because it was challenging, and the feeling of achievement on completion of the task was very satisfying.

Pre-training, only three participants expressed some kind of negative attitude, which was related to the circumstances rather than item-writing in general. Daniel and Logan said they felt worried because they were not sure they would be able to finish the assignment on time, while Josh said he felt bored towards the end of the item-writing assignment.

After the training

Post-training interviews revealed no substantial change in attitudes overall. Participants expressed their positive attitudes to item-writing 53 times in the interviews, which is almost the same as pre-training. *'[I]nteresting', 'liked'/'loved', 'fun', 'enjoy', 'nice'* were terms used on both occasions, but there were also some differences: *'exciting'* and '*challenging'* appeared only pre-training, while '*confident*' was used exclusively post-training. This might be because the initial excitement of doing a novel activity naturally subsided over time, while at least some of the participants felt more confident about item-writing having received training on it. Moreover, none of the participants expressed any negative attitudes to item-writing after the training, compared to three people before it.

4.3.1.3 Use of specifications and example items

The participants were provided with detailed specifications for each item type (see *Appendix 4*), including one example item.

Before the training

From among the 17 participants interviewed pre-training, only about a third mentioned the specifications. Those who did, mostly expressed their attitude to using the specs rather than elaborating on *how* they used the documents. The prevailing attitude was negative: the participants complained that the writing and listening item specifications were long, complex, and difficult to understand, with several participants admitting to not reading the documents carefully:

> *…when I realized the extent of the instructions* [i.e. specifications] *my eyes just glazed over* (Josh)

Notably, participants found the grammar item specifications more helpful, probably because these were shorter and less complex. However, even those who found the specifications helpful could not remember how they used them in item-writing: "[I] *used the specs for A2 and C1* [grammar items], *but I can't remember what I did" (*Ted*).* Only two participants reported more reflective uses: Henry said he repeatedly revised the listening task against the specs, while Logan elaborated on the approach that he used to fully understand the writing prompt specs:

> *I was actually writing them out again in boxes on a piece of paper to make it very clear … I was putting in instructions onto another piece of paper in that process, I was understanding it much quicker as well. It was almost like having two screens for it.*

The dearth of mentions combined with the lack of awareness of how the specifications were used during the item-writing process might be an indication that the participants, who had never written items to specifications prior to the study, did not fully realize the role and importance of this document in item-writing. At the same time, the participants seemed to attach an inflated importance to the example items: a number of participants thought example items alone were sufficient to provide item-writing guidance, with twice as many participants commenting on the sample items compared to the specs:

> *…for all of them I* ***looked*** *at the examples first … it was like* ***glance*** *at the specs and* ***look*** *at your examples and then I started writing them* (Josh).

A '*glance*' suggests very cursory attention, while '*look*' might be interpreted that the item writer paid more careful attention to the example. Stanley expressed his approach even more clearly: "*Basically, I just looked at the example*". Arthur's approach to producing the writing prompt was to *"change it* [the example item] *piece-by-piece to match the specifications asked for"*, something Arthur called a '*retro-fit'*. Participants perceived the example items as models to shape their items, with the words *'model'*, '*template*' and '*example item'* often used as synonyms:

> *There was* ***the example*** *… and I kind of took that as* ***my template*** (Daniel)
>
> *I looked at* ***the model*** *… trying to take that and make mine similar to* ***the model*** (Olivia)

Studying example items was regarded by many participants as the best way of learning to produce a particular item type, with many participants emphasising that they wanted *"to have multiple examples"* (Olivia) and not just one:

> *Like if you've read a hundred detective novels, well then if you're a decent writer you can probably approximate the language and if you're a better novelist you try to write one on your own, but if you've only read one or two of them then it's not so easy* (Josh)

It seems that some of the participants preferred to rely on the example items because they found the specifications too complicated, while others *"felt like it* [the example] *gave* [them] *something to go with*" (Olivia). Both Olivia and Nathan discussed the role of example items for novice item writers; they believed it was natural for a novice to be led by examples, but they also admitted this was not the approach to adopt throughout their item-writing career: *"obviously I won't do that for the rest of my life"* (Olivia).

After the training

Participants' discussion of the specifications was considerably different after the training. Firstly, participants mentioned the specifications more often; secondly, they dwelt on the specifications in more detail. They emphasized reading the specifications carefully and trying to attend to all their aspects: "*I was thinking about the specs, keeping everything… matching everything to the specs"* (Joe). Nathan said that *"**this time** I looked at the specs first of all,"* referring to the fact that before the training he was concentrating on the example item.

Ted described his approach to working with the listening task specifications:

> *It … needs several screens open at once because while you're beginning to write something you have to look back* [at the specifications] *… I had to keep jumping back to the instructions* [i.e. specifications] *then to the text then to the instructions then back to the text and then to the actual items.*

Ted described two methods at once: working from several screens (having the specs on one screen and writing items on the other) and repeatedly referring to the specifications in the process of iterative item-writing. Only Logan discussed working from two screens pre-training; it seems that post-training more participants independently developed the same approach. The iterative approach to working from specifications which Ted had followed was also described by several other participants, especially with regard to the writing prompts: *"I started to write something and then **go back to the specs** to make sure that it complies"* (Henry). Overall, it seems that participants were developing useful approaches for producing items from the specs.

For those participants who talked about the specifications on both occasions, there was a difference in the way they discussed the document. Pre-training, Henry only "*went back to the specs"* after he had written the items and *"realised that I'd missed a couple of things."* Post-training, Henry started the item-writing by familiarising himself with the specs, which took him *"a long time to grasp".* Pre-training, Arthur referred to the specifications only in relation to the example prompt: his approach was to change the example *"piece-by-piece to match the specifications asked for".* Post-training, Arthur focussed on the specifications; he admitted to having problems working with the specifications: *"I'm not very careful reading the specs the first time around"*, so he made an effort to conform to the specifications this time round.

Although Lucy's approach to the specifications did not change – on both occasions she admitted to not being careful with the document, after the training she realised that this might have impacted on her work:

> *I think my problem is that I don't read the instructions* [i.e. specifications] *properly … I'm always rushing doing things, and this is when I'm losing out…*

After the training, participants focussed on the example items much less compared to pre-training, thus only two participants commented on the grammar example item post-training. The attitude to the examples changed too. While pre-training the participants took guidance from the example items much more than from the specifications, post-training the examples played a secondary role. The view of the example item as a 'model' almost disappeared; only two participants, Olivia and Jake, still referred to the example items as models. For example, Jake's approach to producing the writing prompt did not change from before the training: *"I approached it* [the writing prompt] *really by copying the structure of the original email… keeping me very close to the model".* Olivia, however, although she still mentioned '*the model'*, demonstrated a change in approach:

> *…during the course, the way I approached everything including this* [i.e. writing prompts] *was to write something as close to the* ***model*** *as possible, and in this one I actually started by writing something closer to the* ***model*** *and then I thought: 'alright, I don't have to write it about coffee breaks, or anything too close content-wise'…*

It seems that, after the training, Olivia felt more ambitious and deviated from the *'model'* to write a more original item.

4.3.2 Item-type specific findings

This section presents findings specific for each of the three item types produced by participants: grammar items (4.3.2.1), writing prompts (4.3.2.2), and listening tasks (4.3.2.3). Participants' comments on the objectively- and subjectively-scored criteria are discussed in separate sub-sections. Findings on the objectively-scored criteria are organised by criterion. For the subjectively-scored criteria, findings are organised (1) by participant profile and (2) by criterion. Findings from interviews with participants in A-C profiles, which are most interesting for analysis, are discussed separately for each item type. Findings from interviews with profile D participants, where relevant, are presented in criteria-specific sub-sections for each of the three item types, such as 'construct' or 'distractors', but are not discussed separately as a group. With regard to the subjectively-scored criteria, only the criteria that provoked most comments are presented separately for each item type.

4.3.2.1 A2 and C1 grammar items

This section contains grammar item-specific findings from the pre- and post-training interviews. The findings related to the objectively-scored criteria and the ones related to the subjectively-scored criteria are discussed separately.

Comments on the objectively-scored criteria

Mean scores on all objectively-scored criteria of both A2 and C1 grammar items either increased after the training or stayed equally high, with the overall increase in scores being statistically significant. In the interviews, participants discussed four requirements related to the objectively-scored criteria: topic, function, vocabulary frequency, and word-limit.

Topic and function

Pre-training, topic (G8) and function (G9) criteria received lower scores than most other objectively-scored criteria, while 13 participants discussed their choice of topic and/or function for the items in the interviews. Three of those participants scored '0' on both criteria – Daniel, Ted, and Henry. Daniel did not find choosing a topic/function difficult, while Ted said that *"finding a suitable topic … I found a little bit tricky*". Neither of them mentioned using the *Core Inventory*[8]. On the other hand, Henry talked about consulting this document for the topic and the function of his items. It thus appears that the awareness of where to find relevant topics/functions did not guarantee those were selected appropriately. In contrast, all participants who achieved band '2' on both criteria attended to the choice of topic/function and mentioned referring to the *Core Inventory* document. All but one also emphasized that they consulted the document *before* starting to write the items.

After the training, the mean scores on these two criteria increased substantially, with topics and functions discussed much less in the interviews; only five participants volunteered explanations, compared to 13 participants pre-training. It seems that participants were getting

[8] The Core Inventory for General English (British Council – EAQUALS) outlines topics, functions, and grammar exponents for each CEFR level

into the habit of checking the topic/function with the *Core Inventory* and, in contrast to the pre-training, all participants sounded confident with using this document.

Vocabulary frequency

Pre-training, participants were generally able to comply with the vocabulary frequency requirement (G7, A2 M=1.84, C1 M=1.92). Among eight participants who mentioned the requirement in the interviews, seven scored band '2' with all seven talking about using *Lextutor*[9]. Two other interviewees, Lucy and Jake, received low scores on the criterion but did not discuss it in the interviews. This lack of mention might suggest that Lucy and Jake were not aware of this requirement.

After the training, the mean scores for the vocabulary frequency criterion were higher for both A2 and C1 items, with six participants discussing the criterion in the interviews. Five of them just mentioned matter-of-factly that they checked the lexis with *Lextutor*. However, Stanley talked about the requirement at length, saying it was unreasonable:

> *…some really basic words are higher than what you would expect them to be… For example, if you look at K-1 you've good words like 'opportunity' which is a K-1 word. But… 'toilet' is a K-2 word, now I would have thought you'd use 'toilet' long before you'd ever think of using a word like 'opportunity'!*

Stanley pointed to an important distinction between the frequency of vocabulary use by native speakers and the order of vocabulary acquisition by learners. This issue was brought up by different participants (see e.g., the discussion of the input text authenticity in Section 4.3.2.3), which is an indication of the participants' increased understanding of, as well as a reflective attitude to, the specification requirements.

Word-limit

Before the training, most grammar items received band '2' for the word-limit criteria (G1 & G4) and the criteria provoked much fewer mentions compared to the topic and function ones. Three participants talked about the word-limit pre-training, among them Daniel said he found the requirements *"tricky",* especially for the C1 items. James, whose A2 and C1 grammar items scored the lowest on the sum of the objectively-scored criteria, admitted that he did not

[9] Compleat Lexical Tutor website (www.lextutor.ca) contains a VocabProfile tool that matches words of a text to the words of a corpus-based frequency list. The frequency list derived from the British National Corpus (BNC) was used to profile vocabulary for this training course.

understand some of the specification requirements related to the objectively-scored criteria, including the word-limit:

> *I noticed just that the word count, there's a word count in the options, you can see that the way I've written them is incorrect … so obviously that was a mistake by me, just noticed that, so sorry about that.*

Notably, James was able to realise his mistake unprompted by simply going over his items during the interview. This suggests that encouraging item-writer's reflection on the items might help with item-writing skill development.

After the training, still more participants were successful at meeting the word-limit requirements, with only Lucy talking about them in the interview. Both her items received band '2' for the requirements pre-training; post-training, she reported a method that made complying with the requirements easier:

> *I wrote down the number of words required for a stem and for options … I like to have it written down … make sure that I don't go over the word-count.*

Comments on the subjectively-scored criteria

Findings by participant profile

1) Profile A: 'Outliers'

Pre-training, there were three participants whose A2 items scored the lowest on the sum of the subjectively-scored criteria (Logan, Adam, and Olivia), while James produced a low-scoring C1 item.

Their interview responses revealed some commonalities in their pre-training item-writing approach which could explain the low scores: 1) they were guided by the example item more than by the specifications which they hardly mentioned, if at all; 2) they either did not consult the *Core Inventory* for the item construct, or remembered about the document after they had written the items; 3) they mostly talked about the objectively-scored requirements with hardly any mentions of the subjectively-scored ones. The four item writers also struggled to meet three requirements: targeting the construct in items (G16), contextualising the construct in the stem (G11), and creating strong plausible distractors (G12).

The considerably higher scores these participants' items received after the training point to a change in item-writing approach. Indeed, after the training, their discussion of the grammar item production process was very different. Firstly, all of them discussed using item-writing documentation such as the specifications and the *Core Inventory*. Furthermore, the discussion's emphasis was on the subjectively-scored criteria, in particular on targeting the construct and producing strong distractors. The participants admitted that their item-writing process changed following the training. A comparison of Olivia's interviews serves as a good example:

> *First, I looked at the* ***model*** [here and later bold type indicates my emphasis], *looked at the* ***model*** *and then next thing was basically going into the list of* ***topics*** *and* ***functions,*** *so then trying to take that and make mine similar to the* ***model****….* (Olivia, pre-training)

> *I was looking at the* ***core documents****…looking at the* ***sample questions,*** *the* ***topics*** *and the sample questions for 'wh-questions in the past'. I thought back to the work we did earlier in the course and noted again the things that I hadn't understood at the time, so I was trying to limit, keep as much of it as possible in the* ***stems*** *… I'd had problems* ***targeting the constructs*** *during the course. And when I was looking at the wh-questions in the past I was thinking well, is the way to target that by gapping out the wh-word, or is the way to target that by having the wh-word at the beginning and gapping out something else? And I decided to gap out the wh-word. Then for 'if only/regrets' I actually felt like I understood, I may be wrong, I felt I understood how to* ***target that construct,*** *so I just reviewed again the sample sentences in that* ***Core Inventory*** *document and wrote that* (Olivia, post-training)

Olivia's pre-training discussion was much shorter and centred around the 'model' that Olivia was trying to replicate. Her other concern was the choice of the topic and function (objectively-scored requirements). Olivia's post-training discussion was more in-depth and showed more awareness of the item-writing process. Olivia was much less concerned with the objectively-scored criteria; there was no mention of the example item, while Olivia talked about 'the documents,' including *the Core Inventory*, which she used to clarify the item construct. Olivia's main preoccupation after the training was in targeting the intended construct, something that she discussed in great detail both for her A2 and C1 items. Olivia's discussion revealed an increased awareness resulting from the training; however, it also revealed Olivia's doubts about the details of construct targeting – it seems that the training provoked a lot of questions but did not solve all of them.

Although item evaluations suggest that A2 items were more challenging for participants to produce pre-training, this was not the perception of the participants themselves. For example, both Logan and Olivia reported that the A2 item construct was less difficult to target compared to that of the C1 item. After the training, however, these participants acknowledged the difficulties in producing low-level grammar items:

> *…we think 'OK, it's a low-level grammar, so it's going to be very easy,' but it's not* (Logan)
>
> *I found the grammar items – the A2 I think - the lower level grammar items are hard for me, harder for me…* (Olivia)

They were also less certain about the construct of the A2 item post-training (see, e.g., Olivia's discussion of the '*wh-questions in the past*' above). Interestingly, these participants also reported that C1 items were easier to produce following the training. For example, James said post-training: *"This one* [C1 item] *I didn't find particularly difficult",* although his pre-training C1 item scored the lowest.

2) Profile B: 'High-achievers'

There were seven participants whose A2 grammar items, and eight participants whose C1 grammar items, scored very high after the training, while their pre-training items were already of good quality.

Before the training, most of these participants mentioned using the specifications to produce grammar items. However, the example item was equally mentioned, and no effective ways of working with the specifications were reported. The participants also discussed using the *Core Inventory* and *Lextutor*, although, similar to the specifications, the participants were only getting used to working with the tools. For example, Josh explained how he forgot to check the topic and function with the *Core Inventory* so he later had to "*retro-fit".* Participants discussed the objectively-scored criteria more frequently than the subjectively-scored ones, with no mention of distractors. However, even pre-training these participants were aware of the importance of targeting the right construct: Mason spoke about *"making sure I was aware of what the grammar point was",* while Josh was concerned about the fact that he *"didn't even really know what that really was, 'wh-questions in the past'".*

After the training, the participants' attention shifted from the objectively- to the subjectively-scored criteria, in particular to the construct and distractors (see more about this in the sections that follow). The participants reported useful ways of working with the specifications, such as studying the specs thoroughly before starting to write items (Henry). *The Core Inventory* document was mentioned as often as before the training. However, post-training the participants used it not only to check the function and topic like pre-training, but also to get a better understanding of the item construct:

> *…there's examples like you go down to the end of the Core Inventory and there's examples of what's being targeted* (Josh)

It seems that after the training Josh found a way to clarify the construct which he had problems with pre-training. At the same time, the participants started to realise the limitations of *the Core Inventory* and the necessity to 'dig deeper' into the construct to target it successfully in items:

> *I don't think the Core Inventory covers this, they give you sample sentences, they don't sort of go out on a limb in the way you would in a textbook and say what the elements and structure should be focused on ... Obviously, just from the examples that's not enough context* (Mason).

Another difference was that participants reported effective methods of item writing post-training. For example, Henry discussed the iterative process of finding a stem and trying a range of different options to go with it. Mason discussed his way of choosing what exactly he wanted to target in the A2 item:

> *…wh-questions in the past… had a look at the exponents and then decided which of them I wanted to vary and I decided I would look at word order and tense.*

Overall, before the training these participants already demonstrated careful attitude to item-writing, displayed some understanding of the importance of the specifications and item-writing tools. At the same time, their understanding of the construct was limited, and they paid substantially more attention to the objectively-scored specification requirements. Post-training, their attention shifted to the subjectively-scored criteria, where they displayed both deeper and more thoughtful approaches to the construct and distractor issues. Moreover, their item-writing was enhanced by more efficient use of the documentation and tools, as well as more effective ways of item production.

3) *Profile C: Lower scores post-training*

There were ten participants whose post-training A2 grammar items, and 11 participants whose post-training C1 grammar items scored lower on the sum of the subjectively-scored criteria compared to the pre-training ones. In most instances, the post-training total scores were only one point lower.

It seems that at least some of these participants, while producing items post-training, paid particular attention to one specification requirement, which they also extensively discussed in interviews. These participants' post-training items scored '2' on the relevant criterion, while a different criterion, which was not mentioned in the interview, gained a low score. The cases of Liz and Lucas can serve as examples.

In her post-training interview, Liz emphasised the *'distractors are correct as stand-alone'* criterion (G14):

> *The thing that I found hard was to make sure that the options were* ***all grammatically correct in their own little part*** *… to come up with three* ***correct stand-alone options*** *was a bit tricky.*

Liz's A2 item scored '2' on the criterion, while it scored '1' on the stem contextualization criterion (G11) compared to band '2' Liz's pre-training A2 item scored on the same criterion.

Lucas' focus while producing his post-training grammar items was on distractors: *"…so that they were all correct by themselves* [G14 criterion], *but only one of them actually fits in correctly* [G13 criterion]". Lucas' post-training A2 item scored high on both criteria; however, it scored one point lower compared to his pre-training item on '*the key does not stand out from the distractors'* criterion (G15).

Overall, these participants' discussions of the post-training item-writing process were more in-depth and displayed most qualities that characterise the discussions of the *'high-achievers'*. For example, they talked about using item-writing documentation, did not emphasise the example item, and focussed their post-training item-writing discussions around the subjectively-scored criteria requirements. It seems that one thing that distinguished them from *'high-achievers'* was a somewhat skewed attention to some requirements at the expense of others. It seems that, being novices at item-writing, these participants had not yet learnt how to balance their attention equally over all specification requirements.

Findings by criterion

Analysis of the pre-training quantitative data identified two areas of concern: targeting the construct and writing MC options, in particular creating strong and plausible distractors.

1) Construct and item stems

Prior to the training, five participants scored '0' on one or both construct-related criteria: 'the construct is directly targeted in the item' (G16) and 'the stem contextualises the construct well' (G11). None of them mentioned the relevant grammar construct in the interviews. It seems that those participants who failed on a construct-related criterion did not have full awareness of the construct-related requirements. This observation finds support in the fact that those participants who did discuss the construct in the interviews were generally successful in operationalizing it in the items. Several observations can be made about these participants. First, they made sure they fully understood the construct. Whenever they felt unclear about the construct, they sought a clarification using either *the Core Inventory* document, the Internet, or grammar reference books. Having clarified the construct, they gave some thought to operationalizing it in their items. For example, Ted wrote multiple sentence examples and then chose the one that would best target the construct. Moreover, these participants did not think the construct, as embodied in the stem, was separate from the options, but viewed them in synergy:

> *At that point I started playing around with the stem. I say it's the stem, but actually I wasn't thinking about it in terms of that in itself, so I was looking at the whole string* … (Mason)

After the training, the construct received much more emphasis in the interviews compared to pre-training. While before the training participants rarely used the term 'construct' and never used the term 'grammar exponent', after the training the participants used both terms, and they also demonstrated a better construct understanding. For example, Joe discussed the difference between targeting the form and the meaning of a grammar structure:

> *…the challenge of writing grammar items is writing an item that actually targets the meaning of that grammatical structure. I've seen a lot of items that target* [the form]*….but the item doesn't actually target the usage of that structure…*

James' C1 item scored '0' on both construct-related criteria before the training, while he did not provide any comments on the grammar construct in the interview. Post-training, James' C1 item scored '2' on the criteria, and he talked about the need to carefully target the construct: "*Challenge is to make sure you are targeting the grammar rather than anything else*."

Those participants whose items scored high on construct-related criteria both before and after the training, already showed some awareness of the underlying construct before the training started. This understanding seems to have further developed during the training. For instance, Daniel described the process of item creation as a puzzle where all the pieces should come together:

> *it's almost like a kind of a puzzle between getting the topic, function and then allowing some kind of context to generate exactly what the question wants to be targeting, so in this case 'wish/if only/regrets' so it almost feels like a jigsaw puzzle in which you have to get all of the pieces in the correct order.*

Mason discussed how grammar textbooks should be used to clarify details of the construct before writing the items. In the discussion, he also showed awareness of how construct changes with proficiency levels:

> *… different textbooks have their own context, so again B1 have a consensus of exactly what the exponent is … and you introduce the grammar point or structure in different ways depending on what parts you happen to focus on, in terms of ability, the audience – so I don't think there is one answer in terms of exponent.*

2) Options: Key and distractors

While those participants whose items received band '0' on a construct-related criterion pre-training did not mention the construct in the interviews, those participants whose pre-training items had option-related problems often talked about the uncertainty of whether their options were good: *"…my **incorrect options**, I am not so sure anyone would choose them"* (Olivia). Notably, before the training few participants used the term 'distractors', even though it was included in the specifications.

Although aware that producing options was problematic, many participants did not know how to solve the problem. Their understanding of option-production was also somewhat limited. The most often-mentioned criteria were 'distractors are incorrect within the stem' (G13) and 'distractors are grammatically correct as stand-alone words/phrases' (G14). At the same time, two other important criteria – 'distractors are strong and plausible' (G12) and 'the key does not stand out' (G15) were rarely referred to.

The first difference post-training is that participants used the term 'distractor' much more often, while they also seemed to have a clearer understanding of what a good distractor was. If before the training only the most successful participants mentioned some (limited) aspects of option-creation, after the training participants discussed a broader range of requirements, with the requirement for distractors to be strong and plausible discussed most often. Participants said this requirement was the hardest to meet:

> *The biggest challenge in the grammar items was I think plausible distractors. The difficulty so often is writing something that is wrong, but isn't too obviously wrong* (Jake)

Another difficulty was to produce two distractors that were equally strong:

> *I think it is easy to create one option and an option 2, but then there should be option 3, to find the second distractor* (Lucy)

Finally, participants were better able to verbalize their approach to producing item options in the post-training interviews – something that did not occur pre-training:

> *At first, I was thinking 'wouldn't', 'couldn't', 'shouldn't' as a possible, as the three options, but I think that 'couldn't' and 'shouldn't', you could argue that they'd be OK or that they're … at least close enough to be OK without being unfair, so I went with the 'can't' and the 'won't' because I knew that they definitely weren't correct* (Adam, about C1 item)

3) Participants' own grammar knowledge

An important aspect of creating grammar items was participants' own grammar knowledge which was often discussed in the interviews. Pre-training, all participants but one who discussed their grammar knowledge complained that the 'specifics' of grammar were difficult for them as native speakers to understand, even though they had a relevant degree and teaching experience. Among the reported difficulties were grammar structures for different proficiency levels (Logan), grammaticality according to "*grammar books*" vs *"the grammar mistakes that native speakers would make"* (Adam), and unfamiliarity with a particular structure to use in items (Arthur). Before the training, Rose repeatedly said that she was not sure about the construct because of her lack of grammar knowledge, and that she hoped the training would make her *"re-visit a lot of grammar and kind of sharpen up on it"*, thus misunderstanding the training aims.

All participants who discussed their grammar knowledge post-training largely repeated what was said before the training: their uncertainty of (1) grammar structures in relation to proficiency levels and (2) what constitutes 'correct' grammar. Post-training, however, some of them also admitted that writing grammar items was not something they liked:

> *For me the grammar stuff is always tricky … I'm not good with that, so whenever I'm writing grammar items I don't feel very confident with it and it's not the kind of item-writing I'd like to do* (Adam)

4.3.2.2 B2 writing prompts

Similar to the discussion of grammar item-specific findings from the pre- and post-training interviews, the writing prompt-specific findings related to the objectively-scored criteria and to the subjectively-scored criteria are discussed separately.

Comments on the objectively-scored criteria

The writing prompts produced by participants prior to the training received high mean scores on most objectively-scored criteria except proofreading (W5); all these criteria, except proofreading, were also discussed in the pre-training interviews. Compared to pre-training, the prompts produced after the training had higher mean scores for the grammar level (W3), vocabulary frequency (W4), and proofreading (W5) requirements. However, the mean scores for the two word-limit criteria were lower (W1, input message word-limit) or did not increase (W2, overall prompt word-limit) post-training.

Grammar level

Of the five participants who talked about the grammar level requirement pre-training, four were not sure whether they met the specifications (although three of them scored '2' on the criterion) because of lack of understanding of what A1-B1 grammar is. Lucy and Jake reported trying to use *"relatively simple structures"* to ensure the requirement was met. Lucy knew she could have checked the grammar level with *the Core Inventory* but she did not, while Henry did but his prompt still scored '1'. It seems that, although Henry used the document provided, he could not do this effectively. The requests from some participants to be instructed on how to use the specifications and other relevant documentation during the training suggest that this was a broader problem:

> *I wasn't following the specifications well enough, so that's something, that's a skill that needs to be improved* (James)

After the training, the grammar requirement was still perceived as challenging for two reasons: it was difficult *"to present what interests you with quite restricted grammar range"* (Mason) and, similar to the situation before the training, participants were not confident in their own ability to judge the level of grammar structures. For example, Henry expressed this concern both before and after the training. However, post-training his prompt scored '2' on the criterion, which might suggest that the training in using *the Core Inventory* document was helpful for Henry, while his confidence level was still low.

Vocabulary frequency

All those interviewees whose writing prompts scored '2' on the vocabulary frequency criterion reported checking the vocabulary frequency with *Lextutor*. Of those who did not achieve band '2', two - Rose and Lucy – admitted to not having checked the vocabulary frequency with *Lextutor*, while the third person – James – did check the lexis but still scored '0', similar to what happened to Henry with regard to the grammar level requirement. The case of James was discussed in Section 4.3.2.1 with relation to grammar items: he did not understand some specification requirements for the objectively-scored criteria relevant to all item types. After the training, James' writing prompt scored '2' on the vocabulary frequency criterion, with James saying *"*[I] *found things like the lex*[is] *quite easy."* It seems that some novice item writers, as James' and Henry's examples demonstrate, might require training in using item-writing tools and guidelines, while for other novices such training might not be essential. Having been provided with the training, the former item writers are able to meet objectively-scored requirements equally well.

Word-limit

Before the training, all those interviewees whose writing prompts scored band '2' on the word-limit requirements mentioned this requirement in the interviews, while none of those whose prompts received a lower score on this criterion did. Therefore, those participants who did not discuss the requirement might have been unaware of it, which, in turn, led to the lower scores. However, even those whose prompts scored '2' said the requirement was difficult to meet because of the need to include sufficient information in the input message to allow for an appropriate response. Two participants – Olivia and Mason – connected the word-limit to the construct:

> *…you have a task where there's space in order to … express disagreement, explain something and then suggest something, so that within the construct of the text there has to be space in order for the described process to take place* (Mason).

The word-limit for the input message (W1) must have been particularly challenging because its pre-training mean score was quite low (1.68) while the post-training one was still lower (1.64). As different from pre-training, after the training the requirement was also mentioned by those who scored '0' (Olivia) and '1' (Henry). For instance, Henry insisted that he *"checked* *the length and … improved it … trying to keep within the length”*. As the requirement is uncontroversial and easy for an item writer to check him/herself, there might be an underlying reason why it proved challenging even for those participants who were aware of it.

Most participants said that they had initially written a much longer input message and then *“had to cut cut cut”* (Adam). However, reducing the input message length proved difficult, which might suggest that the word-limit requirement is in competition with some other specification requirements. Before the training, Mason and Olivia made a connection between the input message length and the construct, after the training Adam discussed the word-limit in relation to task authenticity:

> *…getting a note on a door to be just sixty words … it doesn’t seem totally realistic... I had it originally saying things like ‘there was an electrical fire, luckily nobody was hurt’, just some little things like that, but it just ended up being 85 words…*

Thus, the input message word-limit requirement might have been problematic for participants not (or not only) because they did not have the ability to write concisely, but because this requirement competes with subjectively-scored requirements such as the construct and task authenticity. It seems that when a (novice) item writer is not skilled enough to comply with several competing requirements, the requirement that is perceived as less important might get superseded with the one that is seen as more important.

Comments on the subjectively-scored criteria

Findings by participant profile

1) ***Profile A: ‘Outliers’***

Before the training, Josh and Ted’s writing prompts scored the lowest on the sum of the subjectively-scored criteria: out of the possible score of 22, Josh’s item scored 16 and Ted’s scored 9, while most other participants’ items scored between 20 and 22.

Table 4-37. Pre-training writing prompt produced by Josh

Instructions to candidates:

You work as the Assistant For Mr Jones. Last Night he Sent You The Following Email:

Input email:

Hi Sherry,

I Need You to Take My Suit TO The Dry Cleaners. My Kid Spilled Coffee On It And I Am Afraid It Has Been Ruined. But First I Need To Know If the Spill can Be Cleaned, How Much It Costs And How Long It will take. I need it for the meeting with our investors on Monday.

Thanks

Steve Jones

Instructions to candidates (continued):

Send an email to Donna at the dry cleaning company asking for the information Mr Jones requested. Write 120 To 150 Words. You Have 20 Minutes.

The writing prompt produced by Josh (*Table 4-37*) scored '0' on the input message genre (W6) and '1' on the input message plausibility (W9) and the construct requirement (W11), among others. Josh's pre-training discussion of the writing prompt concentrated on the fact that he found the specifications difficult to digest (see Section 4.3.1.3), so he tried to *"make my item as much like the example as possible"*. However, it would have been impossible for him – or any other item writer – to fully and correctly deduce all specification requirements from studying one example item. When discussing the prompt construct, Josh mentioned an idea he had but he "*couldn't make it fit*":

> *I was actually thinking … 2 emails, right? Ask the secretary to first write to the dry-cleaning company and then give her that reply, give her another sort of prompt, from here's what the dry cleaning company guy wrote, now you have to report information back to Mr. Jones, so it was like* ***really grilling the reported speech…***

It seems Josh's misconception about the construct stemmed from the fact that he had ignored the specifications, while the example item alone could not have provided him with enough guidance.

Table 4-38. Pre-training writing prompt produced by Ted

Instructions to candidates:

You are a Nigerian Prince. This morning you received the following email:

Input email:

Dear Sir,

We need to inform you that your assets have been suspended by the government. We will release your riches only if you can send us $10,000 by Wednesday of next week. If you do so, your multi-million dollar fortune will again be yours.

Please get in touch if you need any clarification.

Best wishes,
The Prime Minister

Instructions to candidates (continued):

Write an email to everybody in the world asking for help to free your funds. Ask for their bank details and the money that your government has requested. Say how much money you're willing to pay in return for this help. Write 120-150 words. You have 20 minutes.

Ted's writing prompt (*Table 4-38*) scored '0' for the input clarity (W7), suitability for testing (W8), plausibility (W9), and fairness (W16), among others. It seems that choosing an inappropriate scenario was the root cause for the low scores. In the interview, Ted said he was "*trying to find a different or interesting situation*":

> *…it's just the classic email scam. I wasn't really sure what to do. It was a bit light-hearted in the end.*

Finding an interesting situation helped Ted to write the prompt because "*when the situation came it kind of flowed out a bit more easily*". Notably, Ted never mentioned the specifications when discussing the writing prompt. However, unlike Josh, he did not mention the sample item either. It seems that for Ted the creative side of item-writing was what mattered most. It seems that, although Ted and Josh's writing prompts received low scores for different reasons, the problem was the same – a neglect of the specifications while nurturing own ideas of what the construct (Josh) or the scenario (Ted) must be like.

After the training, Josh's prompt total score on the subjectively-scored criteria was '20' (+4 points), while Ted's was '22' (+13 points). Their interview responses revealed some radical changes in their item-writing approach post-training. While pre-training Josh was put off by the complexity of the specifications, after the training he said: *"I conformed to the specs".*

Moreover, Josh's singular focus post-training was on targeting the construct – something that was problematic for him pre-training:

> *...I really tried to make sure that that was* [the right construct] *what I was targeting when I wrote my writing prompt this time.*

However, Josh's interview also revealed a selective approach to the specifications. He invested a lot of effort in targeting the right construct, while the requirement for the input message to be a formal email/notice was overlooked by him both before and after the training, with both prompts scoring '0' on that criterion. It seems that Josh's item-writing skill development happened in the areas he focussed his attention on. It also seems that the time of the training was not sufficient for Josh to master all specification requirements. As a result, although Josh's ability to produce writing prompts improved following the training, his item-writing skills require further development, including the ability to pay equal attention to all specification requirements.

Before the training, and in search of an interesting scenario, Ted produced a prompt unsuitable for testing. After the training, Ted's attitude changed:

> *...this one I had to think a little more. I brainstormed a lot of different ideas before I came to the management one that I came to. The reason that I did it was that* ***it lent itself to a good answer.***

There was a clear change of perspective from creating an interesting prompt to creating a prompt that elicits *"a good answer"*. Overall, Ted's ability to produce writing prompts seems to have improved more drastically than Josh's, with Ted's prompt receiving the maximum score on the sum of the subjectively-scored criteria.

2) Profile B: 'High-achievers'

Seven participants whose pre-training prompts scored quite high (80% or more of the maximum score on the sum of the subjective criteria) produced even better-quality items following the training. For seven more participants, the quality of items stayed equally high pre- and post-training. Analysis of the pre- and post-training discussions of the '*high-achievers*' resulted in several observations. Firstly, most of them mentioned the specifications pre-training and discussed these more extensively after the training. Secondly, they reported some effective ways of producing the writing prompts. Before the training, one participant – Logan – reported an effective way of working with the specifications (see Section 4.3.1.3), while post-training all *'high-achievers'* did so. The most common way was writing iteratively, that is studying the specifications, then producing the first draft or part of it, checking it against the specifications, making changes, for example:

> *I was just going over it and over it and over it* (Henry)
>
> *I got finally a scenario that worked and just iterated on it, making little improvements, reading it again making another improvement…* (Austin)

Post-training, Logan came up with an idea of selecting an input message scenario that would not create problems with the vocabulary frequency requirement. Logan went over vocabulary frequency lists and selected lexis for the scenario before he started to write the input message *"so that I don't have to worry about that while I'm writing it".*

This connection between the objectively-scored requirements and the prompt scenario was realised by some of these participants only after the training (e.g. Henry), while others, such as Logan and Mason, widened their understanding of it. Mason, who discussed the effects of the word-limit on the writing prompt construct pre-training, expanded the discussion post-training to include the grammar and vocabulary requirements, thus seeing all requirements as interconnected. Mason's understanding of the construct was also much clearer and more coherent after the training:

> *…you give them a task which gives the opportunity for a function to be explored in which case this one was expressing disagreement/giving suggestions and decisions. You have to have a prompt which, first of all, gives your students or candidates an opportunity to disagree … and the other thing was to give a clear platform to offer alternative solutions … so those were the main considerations.* [adds later in the interview] *…also that there was enough on a communicative level to stimulate the functions that I wanted to elicit in the writing, so … the disagreement, the suggestions, the decisions.*

Despite obvious item-writing skill development and, consequently, higher prompt evaluations, these participants felt that their item-writing ability could be further developed to make their item-writing more efficient:

> *It took a lot of time … I'm not used to just creating tasks, not yet … I think all the time if I had more of these* [effective ways of item writing] *I'd be able to find shorter ways to do them* [i.e. to produce writing prompts] (Henry)

3) *Profile C: Lower scores post-training*

Eight participants' post-training prompts scored lower on the sum of the subjectively-scored criteria compared to the pre-training ones. The difference in the total scores for these participants' prompts was one or two points only. The input message genre (W6) and input message plausibility (W9) were the criteria which commonly scored lower for these participants after the training. The fact that a number of *'high-achievers'* pre-training saw their writing prompt scores decreased after the training, with the prompts failing on the same specification requirements, might suggest that there were some underlying reasons for the lower scores. Firstly, as reported in the pre-training interviews, these participants tried to closely follow the example item pre-training but, after the training, they decided to adopt a more daring attitude and create a prompt that was considerably different from the example. However, because they did not yet have a fully-developed ability to carefully read and interpret the specifications, they failed to notice some important requirements. Secondly, in an attempt to deviate from the example item, some of them decided to produce a notice rather than a formal email, but they did not seem to have the necessary knowledge of the notice genre conventions, so they failed to operationalize the genre in the prompts. Thirdly, having learnt about some aspects of item-writing during the training, their attention post-training was focussed on these aspects at the expense of other aspects they now perceived as less important. The fact that the participants were novices inexperienced in item-writing might explain their inability to simultaneously pay attention to all specification requirements.

Findings by criterion

1) *Construct*

The writing prompt was designed to assess test-takers' ability to produce a formal transactional email with the purpose of complaining / suggesting solutions / offering advice (W11). The mean score for this criterion was very high (1.92) both before and after the training. It appears that operationalizing the construct in writing prompts did not pose any considerable difficulty to participants. Only two of them, Mason and Olivia, talked about the writing construct in the interviews, both pre- and post-training. They reported paying attention to the construct requirements in the specs and thinking about the intended response in terms of the construct. Post-training, Olivia also reported a way to clarify the meaning of the construct – she *"searched the Core* [Inventory document] *to see for the B2 what might be appropriate".*

2) *Input message-related criteria: genre, plausibility, fairness*

Participants spoke a lot about finding a suitable scenario both pre- and post-training. Pre-training, 12 out of 17 participants discussed the scenario in the interviews. The quantitative analysis revealed that scenario plausibility (W9) was the lowest-scoring of all subjectively-scored criteria pre-training (M=1.6), with plausibility being a major concern for those item writers whose prompts generally scored high on the sum of the subjectively-scored criteria. James, Logan, Lucy, and Mason reported trying to come up with a plausible scenario that would also fit the specifications. For example, Logan thought of messages to residents in a housing estate he lived in, while Mason started by looking through his mailbox to see what messages he had received recently. James made a link between the input message plausibility (W9) and the genre requirement (W6):

> *it does say* ***formal email*** *– so I was just thinking what kind of situation someone would receive a formal mention or email … customer services … struck me as something that would* ***fit those specifications****…*

One might notice a contrast between James' approach and the ones by Ted and Josh discussed above; while James saw a direct link between the specifications and the scenario, Josh and Ted had no such awareness.

Those participants whose items gained a lower score on the plausibility criterion either reported it difficult to comply with the requirement or did not mention the requirement in their interviews. It seems they had other priorities while choosing a writing prompt scenario; for example, Jake's main consideration in choosing a scenario was to elicit a wide range of lexis and grammar.

After the training, the choice of scenario remained a major consideration for participants, with two differences – there was more emphasis on task fairness and situation formality. Twice as many participants talked about creating a culturally unbiased prompt that most test-takers would be able to relate to:

> *…to make* [it] *generic enough for an entire audience across the world without having to have any top-down knowledge of circumstance* (Daniel)

The requirement must have gained prominence during the training following the input on test fairness. Consequently, the mean score on the prompt fairness criterion (W16) was higher post-training.

Only James talked about the formality of the input message (W6) pre-training. Post-training, he also mentioned paying attention to *"the appropriate level of formality".* Four more participants mentioned the formality requirement post-training, showing a realization that formality was an underlying factor in choosing the prompt scenario:

> *…the very first thing that I focused on was the fact that it was formal and so I wanted a context that would elicit that kind of language, so it had to be someone that was unknown* (Joe).

The input message genre requirement (a formal email or a public notice) deserves a more detailed discussion. Five participants did not cope with the requirement pre-training, including Nathan whose pre-training interview contained an interesting detail:

> *It's a case of **sticking closely to the example,** and just thought inspired from there, inspiring from something that I encountered not too long ago, **an announcement for cuts in building** was fresh in my mind and so used that.*

On one hand, Nathan was trying to *"stick to the example"*, which was an email; on the other hand, he wanted to use a situation from *"an announcement".* As a result, Nathan defined the input message as *'a notice'* in the instruction to test-takers but wrote it as an email because he was trying to copy the example's input format. It seems that Nathan did not notice a conflict there. Post-training, Nathan used "a customer services' situation", which he described as an email and framed accordingly. A conflict did not happen, and the prompt received band '2' on the criterion. Nathan's example supports the observation that slavish and unthinking copying of the example item while not paying attention to the specification requirements resulted in lower-quality items by this study's participants.

Despite more awareness of the genre and formality requirements shown in post-training interviews, there was still a number of participants who received lower scores on these criteria. One of those was Josh, whose lack of awareness about the requirement was discussed earlier in this section. However, three other participants emphasised the input message genre in their post-training interviews, but, surprisingly, received a low score. Even more surprisingly, pre-training their prompts gained band '2' on the criterion. A closer analysis revealed a common root to the problem: these were the only three participants who chose to produce a public notice input message post-training, while all three produced an email input message pre-training successfully. The problem for all three was that, while they claimed their input messages were '*public notices*', Joe's was, in fact, *"an appeal from a newspaper editor"*

(Expert Judge 1) published in a local newspaper, Arthur's notice was framed as an email, and Adam's input message had no clear genre characteristics.

Their main reason for choosing the notice genre post-training was to write something they had not done before, for example Adam said:

> *I was looking at the part that says it could be an email, it could be a notice… and* ***I wanted to do something different, something I hadn't done before,*** *so I took the idea of a notice, and I thought what kind of notices are there, and I thought OK the landlord sticks something through your door sort of notice, so that's where I took the idea from.*

It seems that they made a point of not following the example item slavishly; they did not mention the example in the interviews but talked extensively about the specifications. However, possibly because of a limited understanding of the notice genre's features, their post-training prompt scored '0' on the genre requirement.

3) Writing prompt authenticity

Several participants reported paying attention to the prompt authenticity (participants often called it 'naturalness') that encompasses both the characteristics of the input message (W9) and of the intended response (W12). They saw the need to *"…come up with something realistic, … something that you could receive in real life and something that is met with an actual response*" (Lucas). They also considered the genre they had selected from the authenticity point of view: *"I actually started from a position of authenticity … looking at the sort of conventions you'd expect in that kind of text"* (Mason).

The participants predominantly used the term 'authenticity' after the training, and the criterion generated more discussion, compared to before the training. Pre-training, Mason reported prompt authenticity to be a challenge; post-training, he started the item-writing process "*from the most authentic piece of written communication, personally remembering and taking it from there*". Logan reported a similar approach: *"it was really just based on writing an email for some people at work."* Both Logan and Mason scored very high on the sum of the subjectively-scored criteria pre-training, and their post-training gain was 100%. These examples provide support to the observation that high-achieving participants showed awareness of specification requirements even before the training, while after the training they developed ways to deal with item-writing difficulties.

Most participants who discussed the input message plausibility talked about the difficulties in achieving it because of the necessity of balancing it with the objectively-scored requirements such as grammar level, lexical frequency and, most importantly, the word-limit. Interestingly, the intended response plausibility (W12) was not affected by this problem, so the W12 mean score was substantially higher after the training (1.82 pre and 2.0 post), while the W9 mean score was 1.6 and 1.68 respectively, which were the lowest mean scores on both occasions.

4) Instructions

Pre-training, four participants commented on producing an instruction to test-takers, with Emily and James reporting that writing the instruction was the most challenging part of the prompt. The difficulty, in both cases, seems to have stemmed from the item-writers' desire to reproduce the 'model instruction' too literally without considering either the specifications or the difference between their scenario and the example item's:

> *…there were two parts to it in the example, explaining disagreement and suggest alternatives, so I was conscious …there were two points that had to be mentioned to candidates… that was probably the most …the part that took longest regarding the writing* [prompt] (James)

Having "*two parts"* in the instruction was not, in fact, a specification requirement. James and Emily interpreted the example item as a prescriptive format rather than as one of the possibilities of realising specifications in a concrete item. In contrast to James and Emily, those participants who reported attending to the specifications did not discuss the "two parts" of the instruction in their interviews. Among them, Logan mentioned enjoying producing the instruction because:

> *…often we just look at the prompts …and we don't think about instructions because they're often always the same.. so it was quite nice to actually to check those words and make sure they fitted in with what it should have been…*

The instructions attracted fewer mentions after the training. James, who discussed instructions on both occasions, did not mention the example item's instructions after the training, focussing instead on how well his own instructions reflected the writing prompt construct:

> *It's to make sure that you're targeting the right thing and my instructions to candidates were hopefully doing that.*

4.3.2.3 B1 listening tasks

This section presents listening task findings from pre- and post-training interviews and, similar to the previous sections, findings for the objectively-scored and subjectively-scored criteria are discussed separately.

Comments on the objectively-scored criteria

Similar to what was observed for grammar and writing items, mean scores on most objectively-scored criteria of the listening tasks increased after the training (with the exception of the text word-limit criterion), with the overall increase in scores being statistically significant. In the interviews, participants discussed the choice of topic, vocabulary frequency, word-limit, and grammar level requirements.

Topic

Ten participants discussed their topic considerations for the listening task (L3) pre-training. Ideas for the topics came in a variety of ways: "*from imagination*" (Henry), by remembering radio programmes (Logan), from a recently read book (Lucy), or looking through old photos (Stanley). However, it seems that some participants forgot to consult the list of suitable B1 topics in the *Core Inventory* because nine tasks scored '0' on this criterion. The genre requirement (a lecture/presentation/radio programme/news report, L13) might also have been conflated with the topic requirement by the participants untrained in using the specifications – some of them discussed 'the topic' while in fact talking about the genre. Among the considerations for the choice of topic were the topic being broad enough (Daniel), realistic (Henry), and allowing for the text to be informative to support a range of items (Jake). Most participants found choosing a topic difficult and it *"took ages"* (Rose). However, once they arrived at a suitable topic, *"the rest comes a bit easier, because that's the overarching thing"* (Ted).

After the training, the inspirations for topics were similar: they were *"based on personal experience"* (Mason, Daniel), came from searching the web (Lucy, Emily), or after watching a film (Lucas). However, there was one important difference: the majority of participants reported consulting the B1 list of topics in the *Core Inventory.* As a result, the topic criterion mean score was considerably higher post-training (1.76 compared to 1.32 pre-training), with no participant scoring '0'. While choosing the topic, participants took into consideration whether it was suitable *"to give a narrative"* (Lucas) and whether it allowed for a range of information in the text (Austin). Olivia reported an interesting approach to ensure the topic was compatible with the vocabulary frequency requirement:

> *I started off with Lextutor, trying to find something I could write about that was within the appropriate K-level, which actually quickly knocked out a few different ideas.*

It seems that post-training most participants were very clear about the topic requirements – only four out of 25 tasks scored band '1' on the criteria, possibly because the participant forgot about the requirement at the time of writing, as this quote from Adam indicates:

> *I wrote it all and then I realised it didn't really fit the topic … very well, but I put so much time into it that I said "I don't care, I'm just going to pretend it's 'Lifestyles/Describing Events' and I'll hand it in".*

Vocabulary frequency

On average, the pre-training listening tasks scored quite low on the three vocabulary frequency criteria (for the text, item stems, and intended responses). Sixteen interviewees (i.e. all but one) discussed the requirements during the pre-training interviews. Notably, only three said they did not use *Lextutor* to check the vocabulary frequency, which makes one wonder about the low mean scores. One explanation found in the interviews was that participants failed to discriminate between the text (K1-K3) and the stem/response (K1-K2) vocabulary frequency requirements. Only Mason, who scored '2' on all three criteria, said that he *"did the Lextutor stuff on the text first, and then separately Lextutored the responses and stems".* Many participants took K1-K3 as a general requirement, thus failing on the stem and response criteria that required K1-K2 vocabulary. Logan spoke about realising later the difference in the requirements, but he also wondered why it had to be so:

> *if we are allowed to use K-1 to K-3 on the lexical level* [in the text], *why can't we use K-3 words in the stem? … I didn't check the stem lexical level … I think I missed out on that.*

Only seven participants discussed the vocabulary frequency requirements after the training, while the mean scores for the text (L2) and the stems (L8) criteria were statistically significantly higher than the ones pre-training. Post-training, participants did not elaborate on the requirements – they only mentioned that the requirements were complied with. Notably, the mean score for the vocabulary frequency requirement of intended responses (L11) stayed unchanged at 1.6, indicating that it was still difficult for the participants to meet it. It is not clear why this was the case, with none of the participants mentioning vocabulary frequency of responses in their discussion.

Word-limit

The text word-limit (L1) mean score was lower post-training than before it (1.84 and 2.00 respectively), with the difference being statistically significant (*p*=0.046, *r*=0.28). The interviews revealed that, pre-training, participants either did not find the requirement challenging or that they commented on the difficulty of producing a 300-word text, which they thought was too long. The average input text length pre-training was 225 words, with eight texts not reaching 200 words. It seems that many participants' approach was to produce as short a text as possible:

> *I think my text was on the short side, I think it was 189 words … It could have been a bit longer than this, so there would have been a lot* ***more room to make mistakes*** *and to…* ***for inappropriacies to slip in*** (Lucas)

There was no minimum word-limit requirement for the input text (something that, in hindsight, should have been included in the specifications) so, with all pre-training texts being well under 300 words, the mean score for the criterion was 2.0. However, the post-training situation was very different: the average input text length was 278 words with only one text being shorter than 200 words, and four texts were over the word-limit. It seems that before the training participants were struggling to produce long enough texts, while after the training they were struggling to produce texts that would be under 300 words:

> *…in writing the text, I was* ***thinking about the items*** *at the same time and so in knowing about trying to put these things together,* ***within that sort of 300 words sort of thing,*** *I tried to make sure that it was well-structured,* ***there are bits in-between certain items*** … (Logan)

> … ***trying to make sure that it*** [the text] ***will fit the word count*** *and then all the verbal ticks like 'stops and starts' and 'uh's' and colloquial phrases just trying to make it more realistic throughout* (Ted)

It seems that, post-training, participants were more aware of the text authenticity (L14), in-text distractor (L21) and 'no two pieces of targeted information appear too close to each other in the text' (L23) requirements, and discovered it was hard to reconcile these with the word-limit requirement. These novice item writers might not have had enough skill to comply with all requirements, so they might have decided the subjectively-scored ones took precedence. Notably, the same situation was observed for the word-limit requirement for the writing prompt input message (see Section 4.3.2.2).

Grammar level

Very few participants commented on the grammar level of the listening item stems (L9) both before and after the training (three and two participants, respectively). At the same time, the requirement proved the second most challenging pre-training (M=1.2) and the most difficult post-training (M=1.16, i.e. lower than before). Those participants who mentioned the requirement pre-training were successful at meeting it. They were aware of the requirement and created the item stems with the grammar level in mind. Two different participants spoke about the requirement post-training. Of them, Jake scored '2' on both occasions, while Josh scored '0' before and '2' after the training. Josh said he *"weeded out any complex structures … to make it* [the stems] *appropriate for B1 grammatically speaking".*

Comments on the subjectively-scored criteria

Findings by participant profile

1) *Profile A: 'Outliers'*

Pre-training, Alex, Arthur and Josh's listening tasks scored the lowest on the sum of the subjectively-scored criteria. Two of them – Josh and Arthur – gave interviews pre-training.

Table 4-39. Pre-training listening text by Josh

> *hi, I'm pro golfer Mick McMichaels. I'd like to talk about my new favorite product, animal juice. animal juice will give you the energy you need to keep going strong, all day long. animal juice is packed with 87 essential things, and I need every one of them when I'm out there on the course, sweating my face off. What can weasel juice do for you? Just think: after one can, you will be able to do more stupid things faster with more energy. You will have the strength to save your family in the face of a giant bug attack. You will even have the guts to march into your boss's office and demand a raise. What are you waiting for? Call 1-800-ANIMALZ right now, that's animals with a Z, and for a limited time only, you will receive a tin of mouse butter absolutely free! That's right, mouse butter enables you to slip through impossibly small cracks! A case of 25 cans of animal juice and a tin of mouse butter for only $47.77. Call today!*

Problems with their listening tasks were in different areas: Josh's task received low scores on all text-related criteria, while Arthur's task scored low on many item-related criteria. There were also two similarities – both tasks gained low scores on the in-text distractor (L21) and text authenticity (L14) criteria. Josh's interview might help explain why his text (*Table 4-39*) was deemed unacceptable for testing:

> *It was my loosest task to write .., actually, by this point I was quite bored. My text … was a little bit silly because I don't know, what I tend to do when I'm bored… I just wanted to be done with it to be perfectly honest.*

Josh also admitted to not knowing how to approach writing the text because he had never produced a text of this type before:

> *…my only experience with writing … has been fairly creative or analytical, but this is something different, producing a text to be used in a test… I just didn't really feel I had a frame of reference for it and so I sort of tended to rely on my creative side …*

Arthur also managed to correctly identify the challenges he had met when producing the listening task: in the interview, he spoke about text authenticity and distractors. He found the restrictive nature of the specifications incompatible with the authenticity requirement. As regards distractors, he said:

> *Another difficulty I had … is that the specifications indicate that there should be some distractors … for each of the gap-fills, and for some of them the distractor I felt was a little bit more obvious than others.*

Arthur's task scored low on several other subjectively-scored criteria, but Arthur did not dwell on these in the interview. Overall, both Josh's and Arthur's discussion of the listening tasks was brief and limited to one (text for Josh) or two (text authenticity and distractors for Arthur) points they had concerns about. Notably, they did not provide a step-by-step account of how they approached the production of the task; instead, they went straight to the area of concern and based the discussion around it.

After the training, Josh's task scored higher on most text-related criteria. In the interview, Josh did not talk about boredom or text-writing difficulty. Instead, he provided a step-by-step account of his text-production process. However, it seems that Josh's focus post-training was on creating the text and paraphrasing information in the stems (criterion L10). While his post-training listening task scores were high on the corresponding criteria, his other area of weakness, distractors, remained out of focus, with his pre- and post-training listening tasks scoring '0' on the criterion. It should be noted that a similar situation was observed with Josh's writing prompt, as discussed in Section 4.3.2.2. It seems that Josh tended to focus on a limited number of requirements for improvement at a time.

Arthur’s post-training task received higher scores on the distractor and text authenticity criteria, although there was still space for improvement. Arthur’s post-training interview, similar to pre-training, was exclusively about the two requirements, with Arthur saying they were still difficult to comply with. Notably, both Arthur’s listening tasks scored low on the clarity of the stems (L24), while Arthur did not elaborate on the criterion in either interview.

2) *Profile B: ‘High-achievers’*

Pre-training, the interview analysis revealed a lot of similarities in the approach of those participants whose listening tasks scored high on the sum of the subjectively-scored criteria. At the same time, their discussions differed substantially from Josh and Arthur’s, who produced ‘outlier’ listening tasks. The high-achieving participants said that the listening task was the most difficult and took them the longest. However, instead of complaining or feeling overwhelmed, they tried to develop an approach that would work for them. There are striking similarities between the profile B participants in their approach to producing listening tasks. All of them reported writing the text and the items simultaneously, instead of first producing a text and then items for it, or vice versa, and emphasized that producing the listening task was an iterative process:

> *I tried to think of them at the same time … look at what the stems would be and then try and fit them into the monologue* (James)

They realized that two elements came into producing listening tasks – creativity and attention to detail. They enjoyed the creative part and accepted the necessity of complying with the detailed set of specifications. It seems that all listening tasks that scored high on the sum of the subjectively-scored criteria were produced using a similar approach, while each task that scored low had its own problems.

Post-training, 11 of these participants either produced a listening task of an equally high quality (Nathan) or wrote still better tasks. The post-training interviews revealed that these participants retained most aspects of their item-writing approach, having also improved on it. Similar to pre-training, they produced the text and items simultaneously and interactively:

> *It was a much more back and forth when coming up with the listening text and the questions* (Daniel)
>
> *I tried to do them concurrently* (James)

Nathan described how he developed his approach further:

> *…once I got the idea of an old man talking about his town…I brainstormed things that possibly he could talk about that might be reasonable in this context … so that initial brainstorming for ideas took a little bit of time.*

Introducing a pre-writing brainstorming stage appears to be a more advanced way to writing items, something that no participant reported before the training and only some *'high-achievers'* described after it.

These participants' item-writing approach both pre- and post-training was characterised with attention to all aspects of the specifications and an iterative item-writing process. However, some new aspects were revealed in the post-training interviews. Namely, the participants tried to incorporate what they had learnt during the training into their item-writing:

> *I was again reviewing what we'd done in our groups, and what other participants had commented on when we did our listening tasks* (James)

They also reviewed their pre-training listening tasks in order to reflect on their weaknesses and make sure they wrote a better task:

> *I think one issue I had with my last listening was that there were too many numbers, I think I had too many questions to ask about numbers … so I wanted to do a bit more variety this time* (Nathan)

It should be noted, though, that none of the participants managed to gain the maximum score on the sum of the subjectively-scored criteria before the training, while after it only one participant – Matthew – did so. Unfortunately, Matthew did not consent to be interviewed on either occasion, so no qualitative observations can be made regarding his item-writing approach.

3) Profile C: Lower scores post-training

Eight participants whose listening tasks scored high on the sum of the subjectively-scored criteria before the training received lower scores for their post-training tasks. In most cases, the score difference was not large. Mason's listening task production process post-training was similar to what he did before, that is he produced the text and the items simultaneously. However, if before the training he looked at the task from multiple perspectives, post-training his single focus was on producing distractors. He was particularly concerned with making the distractors more authentic-sounding:

> *… looking again and sort of focusing on distractors and again without it affecting the flow of the speech, you know with distractors it can make everything quite convoluted…*

It seems that the concept of text authenticity was new to Mason and he developed a somewhat limited understanding of it. For him, text authenticity was predominantly about the authenticity of the situation:

> *it's based on personal experience again, it's something I've been to, what is it, an open day for sailing? I've been there, done it – so it gets more likely to be authentic.*

The speech authenticity seems to have been of secondary importance to him; therefore, he made no effort to make the language of the input text sound more spoken-like (see a discussion of text authenticity later in this section). Consequently, Mason's text authenticity score stayed at '1' post-training. Moreover, because of his sole focus on distractors, Mason must have overlooked the construct requirements, with his post-training task scoring lower on two construct-related criteria compared to the pre-training one. In his interviews, Mason did not discuss the listening task construct post-training, while he mentioned it before.

Logan's overall item-writing approach, similar to Mason's, did not change after the training. However, following the training he was much more focussed on the objectively-scored requirements. His post-training task score on the sum of the objectively-scored criteria was higher by five points, compared to pre-training. Logan also paid equally close attention, both before and after the training, to the authenticity of the input text, and his tasks scored '2' both times on the relevant criterion. However, Logan did not discuss distractors on either occasion; his discussion post-training predominantly focussed on the text rather than on the items. This might be why Logan's task scored '2' on all text-related criteria but the distractor-related score (L21) was lower post-training, while the score on the clarity of items stayed the same at '1'. It is somewhat surprising, though, that Logan was able to comply with the distractor-related requirement pre-training, given he did not discuss the requirement in his interview.

Lucas and Emily's post-training listening tasks had higher scores for the input text's authenticity, but the scores were lower or stayed low on a range of other criteria. Comparative analysis of the interviews revealed that both Emily and Lucas paid increased attention to input text authenticity after the training. While pre-training Emily found a written text online and slightly edited it, after the training she came up with an elaborate procedure to ensure text authenticity:

> *I was looking at reviews on technology on YouTube … and then I wrote some prompts based on a YouTube clip … I just wrote a few bullet points from the video because obviously the video is scripted and then I recorded myself talking about the same topics.*

Emily also tried to ensure the authenticity of items:

> Interviewer: *And when you were writing the items, what were you paying attention to?*
>
> Emily: *I think it was* ***the main information*** *in the text... If I really wanted to know about this piece of technology what would I want to know about it? What would I be listening for when in a real situation?*

It would probably be natural for real-life listeners to concentrate on the gist of the text Emily created. However, Emily overlooked the fact that the task construct, as defined in the specifications, was "to test the ability to locate and record specific information". Therefore, while Emily's text scored high on authenticity, the score was low on the construct requirement. Another problem for Emily was distractors – her tasks scored '0' on the criterion on both occasions. Emily did not make any mention of distractors in her pre-training interview, while post-training she said:

> *…distractors as well, I don't think it's always very clear…I don't think I've got it completely clear in my mind what a distractor is in a listening text …* ***maybe that comes with experience*** *and comes from seeing other items.*

It seems Emily's item-writing skill developed in the area she focussed on but stagnated in the area she did not reference. However, her latter comment seems to suggest that, unlike pre-training, she started to develop some awareness of distractors; however, she felt it was something she would only be able to develop with further practice. Emily may have selected authenticity as the area for her immediate attention and put distractors off for later. Emily's case seems similar to Josh and Arthur's, whose incremental skill development was discussed earlier in this section. However, Emily's case is complicated by one circumstance: it appears that her exclusive focus on authenticity not only delayed skill development in other areas, but also caused a decrease in quality on a construct-related criterion. One explanation, as discussed above for Logan, might be that novice item writers need to pay conscious attention to all specification requirements and, if producing an item involves a large set of specifications, it becomes more challenging, especially when there are several interconnected specification requirements competing for the item-writer's attention. This was discussed in the writing prompt section regarding the input message word-limit versus construct/authenticity (3.3.2) and will also be discussed later in this section.

Similar to Emily, Lucas failed to notice some specification requirements in concentrating too much on text authenticity. He wanted to make his post-training input text sound "*very colloquial… very normal, a normal person speaking."* He wrote a dialogue, probably because dialogues often sound more colloquial than monologues, having overlooked the input message genre requirement (L13). Moreover, there was another change to Lucas' post-training approach. Before the training, he said:

> *I decided what I wanted the listening to be about, what topic, then I started on the stems.* ***I wrote the stems first and then the text****.*

Post-training, probably because of his focus on the text, he reversed the order: he developed the text first and *"made up*" items after that. Lucas' pre-training interview discussed both items and the text, while his post-training interview was predominantly about the text. It seems that this change in approach was not beneficial for the task quality, because Lucas' task received lower scores on four item-related criteria. It is worth noting that the approach that worked for all *'high-achievers'* both pre- and post-training (see earlier in this section) was to produce the text and items simultaneously and iteratively. However, it seems that creating items first and then putting the text into words worked better than the other way around. Stanley's case supports this observation: pre-training, he also created items first but reversed the order post-training. Two problems followed, as Stanley described in his interview: firstly, large stretches of the text did not allow for item generation, and secondly, the text did not contain information that could serve to distract. Stanley had to re-write the text, removing and adding parts of it. Luckily, Stanley became aware of the problem and addressed it while producing the task, so the task scores were not affected. However, his post-training task took much longer to produce.

Findings by criterion

Above, I analysed three distinct participant profiles by comparing their pre- and post-training item-writing performance. The qualitative analysis also revealed several aspects of the specifications that received participants' attention. Among them were the construct, the input text genre and authenticity, item stems, and in-text distractors.

1) *Construct*

The construct-related criteria (L19 & L20) generally scored high pre-training (M= 1.88 & 1.92 respectively), with only two participants mentioning the construct in passing. Four participants discussed the construct post-training, the number still being quite low. However, this time the participants demonstrated better awareness that the listening task was supposed to test the ability to locate and record specific information from a text. For example, Joe said *"my main focus was creating a text where the candidate would actually have to focus on listening carefully because it is listening for detail, for specific detail*". Josh, who was confused pre-training as to what listening subskills the items were supposed to target, post-training said *"Well, it obviously had to be specific information"*.

Four listening tasks received lower scores on the construct requirement post-training compared to the pre-training ones - those of Lucas, Emily, Chloe, and Mason. Chloe was not interviewed on either occasion, while the other three participants' cases were discussed earlier in this section so I will only briefly summarise them here. Emily's main focus post-training was on making not only the text but also the items authentic-sounding, which seems to have had a trade-off for the construct requirement - the items targeted the gist of the message and not concrete detail. Lucas paid more attention to the items pre-training, while post-training his attention shifted to producing an authentic-sounding text. He also changed the approach from producing items first to producing the text first. Mason's main preoccupation post-training was in producing authentic-sounding distractors with other item-related criteria not mentioned in his interview. It seems that all three participants shifted their item-writing focus to various aspects of authenticity post-training. Overall, the text authenticity criterion enjoyed a large and statistically significant increase in scores across the whole cohort after the training; however, at least in part this seems to have happened at the expense of several related requirements. The effect of such inter-relatedness was not unique to this situation and has been discussed on several occasions throughout this chapter.

2) *Input text genre*

The specifications identified the input text genre (L13) as a monologue (such as a lecture, public announcement, or a radio programme). Those participants who received band '2' on the criterion were aware of the genre requirement:

> *I liked the idea that it was a monologue - well actually I think it was a requirement – a monologue – so it was the interviewer kind of setting up the background, the top down*

> *knowledge for the students, the schemata, then it goes straight into the monologue…* (Daniel, pre-training)

Those participants whose tasks scored low on the genre criterion overlooked or misunderstood the genre requirement. For example, pre-training, Henry was convinced he had to produce a dialogue: *"[I] tried to write out the dialogue first"*; Josh called the input text "*text*" throughout. This lack of genre awareness led to Josh's text lacking genre characteristics, so it scored '1' on the criterion. Lucas, whose task scored '2' on the criterion pre-training, forgot about the genre after it, so his text was a radio programme dialogue and scored '0' on the criterion. Notably, Lucas realised the mistake, but only after he submitted the post-training item-writing assignment, so he even wrote an email to the tutors to explain the problem:

> *It wasn't really a monologue as I told you in the email, it was a presenter who asks a couple of questions throughout so that doesn't quite qualify as a monologue I suppose.*

Generally, though, the genre requirement saw higher scores after the training (M=1.8 compared to 1.68 pre-training), and the key to complying with the genre requirement seems to be item-writers' awareness of it.

3) *Input text authenticity*

Only two listening tasks – by Logan and Ryan – achieved band '2' on the text authenticity criterion before the training. Logan's approach to creating the text might be the reason for the high authenticity score:

> *I walked around the hotel room, so having this conversation in my head on how it would actually sound … doing the speaking myself … how does this actually sound rather than how does this look on paper.*

Pre-training, text authenticity was actively discussed in the interviews, although the term 'authenticity' was rarely used. Some participants admitted that producing an authentic-sounding text was difficult for them because they were not sure what the authenticity requirement encompassed. Some participants also did not perceive the authenticity requirement as important. As Nathan put it, *"I thought this is a listening task, not supposed to be completely natural sounding really".* That might be why, although the participants were aware of the requirement, they did not make an effort to comply with it:

> *I didn't really have time to go on the internet to focus on an actual 'real-life' listening* (Lucas)

However, Henry reported paying special attention to text authenticity while producing the listening task. He started with choosing a real-life scenario, and then used his creativity to produce a text that would sound authentic *"keeping it* [the input text] *authentic, it can't be too stilted, it can't be as though somebody were reading it".* Despite the efforts, Henry's input text scored '0' on the authenticity requirement: an inauthentic genre – a 242-word turn in an informal conversation - was selected for the monologue *("a student telling his friend about his favorite water park").* Notably, Henry himself perceived his text as a dialogue saying he *"tried to write out* ***the dialogue*** *first*". This failure to identify what was a suitable monologue genre led to two band '0' scores – on the text genre (L13) and text authenticity (L14) criteria.

After the training, the scores on the input text authenticity criterion were statistically significantly higher than the pre-training ones. There were also more mentions of text authenticity in the interviews, and the term 'authentic/authenticity' was used twice as often. While pre-training, participants either did not take the authenticity requirement seriously or did not know how to make their text sound more authentic, post-training most participants reported paying a lot of attention to the input text authenticity and employing various techniques that they learnt during the training. Repetitions, ellipsis, corrections, afterthoughts, redundancies, pauses, false starts, hesitations, fillers, and asides were mentioned as spoken language features that participants used to make their input texts sound more authentic. Among grammatical features, simple short sentences, colloquial grammar, grammatical errors typical of spontaneous speech, elliptic sentence form, and simple conjunctions were mentioned. Participants also discussed the use of colloquial language and emphatic language:

> *…the sentences are not really full sentences, there's a lot of redundancy like 'yeah, like I said', 'yeah I dunno' 'bit of a shame really', these little phrases that English people throw about* (Lucas)

Four participants reported recording the text from content points and then transcribing it, a technique sometimes called 'semi-scripting' (Buck, 2001). Among those four, three gained the maximum score on the text authenticity criterion. Interestingly, two other participants reported attempting the technique but then abandoning it. Mason said that "*It didn't … work, I got a little bit self-conscious*". Stanley attempted the recording without planning text content first. After he had transcribed the text, he discovered that the text could not support items:

> *It … made the questions too confusing in that the answer could have been this or it could have been something else, so it wasn't as clear cut as I thought it could have been. There were too many variable answers...*

A variation on the recording technique emerged during interviews, whereby participants reported vocalising the text from content points without recording it. Among the three participants who used the vocalisation technique, two gained the maximum score on text authenticity. Interestingly, Logan seemed to have arrived at the technique prior to the training as he reported the same approach then, both times gaining the maximum text authenticity score. Notably, none of the participants reported using a transcription of a real-life audio file. However, Emily reported a hybrid of the two techniques: she used an authentic YouTube videoclip to base content points on, and then recorded herself speaking from the content points. Her text received the maximum score on authenticity.

Most participants, however, tried to increase their text's authenticity by inserting spoken language features into pre-written texts, for example Henry "*re-edited in some bits and pieces that I thought would make seem as I had achieved the out-of-mouth".* This approach might have been so widely-used because of its seeming ease, but of the nine participants who employed it, only one gained the maximum score on text authenticity.

Most participants reported that producing authentic texts was a challenge because authenticity had to be reconciled with other specification requirements. Among those were vocabulary frequency, grammatical level, and the need for distractors:

> *…you have this tension on the one hand of producing something that is like native speech whereas on the other hand being incredibly restricted on what you can say in terms of lexis, in terms of grammar, so it's very difficult to make it authentic-sounding listening* (Josh)

The distractor requirement seems to have been the biggest challenge of all, with participants struggling to incorporate distractors in the text without them standing out as artificial:

> *I'm trying to figure out how to incorporate distractors in the text in a way that sounded authentic because often times in our normal speech we speak more deliberately* (Arthur)

4) Item-related criteria

The specifications contained several requirements for items, including that the items should follow the order of the text (L22), should be distributed evenly throughout the text (L23), and should be clearly formulated (L24). The first requirement (L22) had the mean score of 2.0 on both occasions, with no participant commenting on it before or after the training. It seems the requirement was perceived as obvious, easy, and not requiring of comment. The L23 mean score was higher after the training (1.96 compared to 1.84 before it), with only one participant commenting on the requirement on each occasion:

> *…whether the candidate would have enough time to answer the question and then focus back in on the audio* (Daniel, pre-training)

> *I realised I can't have two questions from the same sentence and so I had to change one or two* (Adam, post-training)

Daniel scored '2' on all item-related criteria pre- and post-training, while Adam's L23 score was higher after the training, probably because post-training he was better aware of the requirement.

In contrast, the 'clarity of items' (L24) criterion posed a lot of difficulty to the participants pre-training (M=1.28). Unfortunately, none of the three participants who scored '0' on the criterion pre-training gave an interview. However, seven of those who scored band '1' dwelt on their approach to producing listening task items, in particular in relation to clarity of items. The interviews revealed an interesting difference between these participants' approach and the approach of the five *'high-achievers'* discussed earlier in this section. While the high-achievers wrote the text and the items simultaneously, these participants wrote the two components of the task separately. Adam, Emily, Henry, and Lucy reported writing the text first and then coming up with items for it. Jake and Stanley used a different approach: they first produced items and then a text to go with them, something they called *"working backwards":*

> *I did the items probably before I did the text… I sort of* ***wrote it backwards*** *…* (Stanley)

It seems that producing items and the text separately, whatever comes first, resulted in less clear items pre-training, compared to when the text and the items were produced simultaneously and interactively:

> *…writing a little bit, thinking about what the stem could be … realizing that the stem could be a bit too complicated, then going back into … the text, and then changing it and then thinking about the logical organization of questions about what someone in an interview would talk about.* (Daniel)

The ‘clarity of items’ criterion’s mean score was only marginally higher post-training (M=1.36), with four participants talking about it in interviews. Notably, if pre-training the clarity of items was explicitly mentioned only by high-achievers, post-training it was mostly discussed by the participants whose tasks scored ‘1’ on the criterion. It seems that they had gained an understanding of the requirement, which they wanted to discuss, but did not yet have the ability to realise it in items. Stanley said:

> *…the answer could have been this or it could have been something else, so it wasn’t as clear cut as I thought it could have been.*

Stanley made changes to the task to clarify the stems, but this must not have been enough as his task still scored ‘1’ on the criterion.

5) In-text distractors

By far the most frequently discussed consideration for participants in producing listening task items was distractors. Pre-training, all participants, including those whose tasks gained band ‘2’ on the distractor criterion (L21), reported that creating these was difficult. Some of the participants whose tasks scored low on the criterion reported not having noticed the distractor requirement initially; they then had to add distractors to the text later, which might have made their job more difficult and the distractors less successful:

> *…trying to come up with some distractors was possibly the hardest part, because at first I just sort of wrote it* [the text], *and then I noticed that distractors were needed, so then I had to tweak it* (Adam)

Some of them also overlooked the requirement for each item to have an in-text distractor. For example, Lucas said he *“made sure … there were* ***a couple*** *of distractors*”, that is distractors for only some of the items.

Those item writers whose tasks scored ‘2’ on the distractor criterion interpreted the distractor requirement correctly: “*I had one distractor for each of the items” (*Olivia). They were also aware of the requirement before they produced the text, and they worked to incorporate distractors *while* (and not *after*) writing it. For example, Nathan said he *“tried to throw some of the distractors in there”* when creating the text, while Olivia said:

> *As I wrote the text I was thinking about distractors … so every time I was thinking that this was a potential question, I put in a distractor for it.*

These participants realised the necessity of making the distractors *"fit and … not too obvious"* (James). One way of introducing distractors naturally into the text was choosing a situation and targeting the information that would lend itself well to creating distractors. For example, Jake took the distractor requirement into account when choosing the text scenario:

> *There's a possibility of somebody getting confused in that situation, and that actually gives you some scope to write distractors.*

The requirement for each item to have a strong in-text distractor was still challenging for the participants after the training. In fact, the mean score was 1.28 pre-training and 1.2 after it. Analysis of individual scores revealed that nine participants created high-quality distractors both before and after the training; six participants wrote better distractors post-training; three participants failed on the distractor requirement on both occasions; while seven participants' scores on the distractor-related criterion were lower post-training. I analysed available interviews from participants in each group to look for any patterns that might explain this trend.

Most participants who created high-quality distractors before and after the training discussed distractors in the interviews. They were aware of the necessity to have distractors, and they fully understood the requirement: *"everything has to have a distractor"* (Austin). Secondly, they were clear about the characteristics of strong distractors: convincing, appropriate, not overly obvious, and realistic. Pre-training, these participants described some approaches to creating strong distractors, as discussed above. Post-training, they demonstrated further increase in awareness in this regard. For example, Jake dwelt on differences in creating distractors for MCQ and gap-fill items*: "you're not writing distractors but you need to put distractions into the text, so you need to think of things that are not too obvious"*, while Mason talked about the need to balance two competing requirements – text authenticity and distractors for each item.

For those participants who produced better distractors after the training a change in approach was observed. Pre-training, they did not elaborate on the way they produced distractors, they also overlooked the need to have a distractor for each item. Post-training, their discussion of distractors was more elaborate and concrete:

> *I tried to figure out how many items I needed to have and for each item there would have to be one key and one distractor. I tried to have parts of information that would 'double' … one would be the key and one would be the distractor* (Arthur)

Among the three participants who failed on the distractor-related requirement both before and after the training, Josh and Emily were interviewed on both occasions. Their tasks seem to have scored low on the criterion for different reasons. Josh did not talk about distractors either pre- or post-training. His listening task was of much higher quality after the training, as discussed earlier in this section, but the gain happened in areas other than distractors. It seems that Josh simply overlooked this requirement. Emily, however, became aware of the distractor requirement after the training but, as discussed earlier, she was not sure what constituted a good in-text distractor and was hoping to develop the ability to produce distractors later into her item-writing career.

Only two participants whose post-training tasks scored lower on the distractor-related criterion offered comments in both interviews. It seems that Henry was not completely confident about including distractors pre-training. Post-training, he *"added in **a few distractors** into the text just a little bit so there's a higher level of difficulty."* Probably, the fact that he added only *"a few distractors"* can explain the score of '1'. As for Olivia, she was clear about the distractor requirement pre-training, as discussed above. Post-training, she sounded more confused:

> *I'd try to put some things that would work as distractors like you can see that for the movie theatre I'd put in **maybe in an unfair way, I hope not in an unfair way** … So I had all these numbers in there which I thought **might make it more difficult for them, I hope not too difficult …***

Notably, both Henry and Olivia produced higher-quality post-training listening tasks overall, with high scores on several criteria that were problematic before the training. However, the distractors were not one of them. It might be that the participants shifted their attention to the areas of weakness, which caused less attention to be paid to distractors. Paradoxically, the learning that happened during the training might also have caused a drop in distractor quality. For example, having learnt about test fairness and level-appropriate challenge, Olivia was very concerned with them post-training, which is evident from the quote above. It seems that more knowledge resulted in more uncertainty, and a longer time might be needed for Olivia to transfer her theoretical knowledge into improved ability to create in-text distractors.

4.3.3 Summary of the qualitative findings

Following their first experience producing items to specifications for the pre-training assignment, all participants reported that item-writing was difficult, with listening tasks perceived as most difficult to produce, while writing prompts posed the least challenge. Remarkably, after the training only a few participants thought item-writing was easier for them, while the rest said it was still difficult because it is generally a challenging activity and also because, having learnt about item-writing during the training, the participants realised there was a lot more to it than they had thought at the beginning. Participants also believed that the induction training was only the first step in their development as item writers, hoping that with time and experience producing test items would become easier or, at least, less time-consuming.

Despite the difficulty, most participants enjoyed item-writing both before and after the training. The activity was seen as stimulating, intellectually challenging, and providing variety to the participants' work life. Post training, participants also reported increased confidence when writing items which, however, did not necessarily lead to the item writing being easier.

Before the training, participants mentioned example items substantially more often than specifications when discussing grammar and writing items. Participants saw example items as specifications in their own right instead of being just part of the specifications. Participants also perceived the example items as a 'model' to imitate. When participants found the specifications difficult to understand, they resorted to copying the example item. Sometimes, such over-reliance on the example resulted in misinterpreting the specifications. For instance, two participants spoke about the writing prompt instruction having to consist of two parts because the example item had this structure; however, this was not a specification requirement. It should be noted that the example item for the listening task did not have as much influence on participants as the grammar and writing example items did, possibly because participants found the listening task much more difficult to imitate without it being too obviously 'a copy'. Therefore, those participants who found the listening specifications too challenging to understand tended to just ignore the specifications and 'guess' the requirements. Some items received low scores as a result of this approach due to incorrect assumptions about some aspects of the specifications.

Participants' attitudes towards working with the specifications changed after the training – they realised the importance of specifications in item-writing and the necessity of studying them thoroughly, something that cannot be replaced with copying example items. Although some participants still mentioned the complexity of specifications, particularly for the listening task, the difficulties did not deter participants from using the specifications to produce items. Many participants reported ways of working from specifications, for example, using two screens to look at the specifications and write items at the same time, keeping in mind all aspects of the specifications, writing out main points from the specifications for easy recollection, and referring to the specifications repeatedly while producing items.

Example items, which were mentioned much more often than specifications pre-training, received substantially less emphasis after it. Moreover, example items were no longer seen as models to copy. Some participants, led by the desire to write something more original, produced writing prompts that were substantially different from the example. Unfortunately, such experiments sometimes lead to lower-quality items because the participants were unable to fully conform to a specification requirement which was not exemplified in the sample (especially in terms of the input text genre for the writing and listening items).

Findings related to the objectively-scored criteria for all item types seem to suggest that the following were sufficient for most participants to produce items that met most/all of the objectively-scored criteria: 1) being aware of the objectively-scored requirements, 2) studying the requirements thoroughly before producing an item, 3) referring to the specifications repeatedly while writing the item, and 4) revising the item. However, some novice item writers seemed to need more guidance than others in interpreting objectively-scored specification requirements. Moreover, complying with objectively-scored specification requirements often required the use of additional documentation and/or online tools. The findings indicate that simply including a document in the item-writing pack or providing the link to an online tool might not be enough and training might be required. For example, several participants who reported using the *Core Inventory* and *Lextutor* still failed to comply with the relevant requirements. After the training, all item writers demonstrated much better familiarity with the item-writing documentation and tools, as well as confidence in using them. No misunderstandings were reported in applying the documentation or using the tools post-training.

The majority of item writers whose pre-training items received low scores on the objectively-scored criteria wrote much higher quality items in terms of the objectively-scored requirements after the training. Notably, the increase in quality on objectively-scored criteria coincided with a decrease in mention of these criteria in the interviews. However, the number of mentions also depended on the item type: objectively-scored criteria were talked about much less in relation to grammar items but more for listening items. The difference seems to relate to the complexity of specifications: listening task specifications were perceived as substantially more complex, so compliance with all requirements, including objectively-scored, required more conscious attention on the part of the participants.

Despite generally higher scores on the objectively-scored criteria after the training, three objectively-scored criteria (one for the writing prompts and two for the listening tasks) had lower mean scores. Analysis revealed that these criteria were competing with one or more subjectively-scored requirements. For instance, the writing prompt's input message word-limit was related to the construct and plausibility requirements, while the listening text word-limit was connected to the text authenticity and item distractor requirements. It seems that, pre-training, participants focussed on meeting the objectively-scored criteria, possibly because the criteria were more obvious and easier to comply with. After the training, with the increase in awareness about such concepts as construct, test authenticity and distractors, participants' attention shifted to the subjectively-scored requirements; however, because the two requirements were related, the objectively-scored requirements saw a decrease in attention and, consequently, in scores. The drop in scores on the objectively-scored criteria, however, generally did not occur for items produced by profile B participants, so-called *'high-achievers'*. For other participants, however, meeting several competing specification requirements could prove excessively challenging even after the training.

Gain ratio analysis of the total scores on the subjectively-scored criteria revealed four participant profiles, three of which, as the most interesting ones for analysis, were discussed in this section. Analysis of the interviews for each profile revealed some similarities in the participants' approach to item-writing.

Profile A participants produced the lowest quality items on the sum of the subjectively-scored criteria before the training but invariably wrote items of markedly higher quality after it. Some reasons for their items' low-quality pre-training could be discerned from their interviews. For instance, they perceived item-writing and especially using specifications as difficult, and their way of overcoming the difficulty was in neglecting the specifications and copying the example item instead. When this was not feasible, such as for the listening task, they made guesses about some of the requirements. They nurtured their own ideas, probably inspired by their previous teaching or testing experience, of what an item should be like and, because of the lack of attention to the specifications, they could not check those ideas against specification requirements. Overwhelmingly, they paid more attention to objectively-scored requirements while neglecting subjectively-scored ones, or they paid attention to some of the subjectively-scored requirements only. After the training, their approach to producing items underwent a radical change – something they realised themselves by contrasting what they did before the training with what they did after. They reported no negative feelings towards item-writing and using the specifications. They stopped exclusively relying on example items; instead, they attended to the specifications. They demonstrated more awareness of the item-writing process and reported some item-writing approaches they had not used pre-training. They started paying more attention to the subjectively-scored criteria. It is important to note, though, that none of the *'outliers'* items, although markedly better than the ones produced pre-training, received the maximum total score – there were still some areas of weakness in each item.

Profile B participants, or *'high-achievers',* were those participants who produced good quality items before the training and whose post-training items were of similar or higher quality. *'High-achievers'*, although admitting pre-training that item-writing posed considerable difficulty for them, did not concentrate on the difficulties but looked for ways to overcome them. They also seem to have arrived at an effective way of producing a particular item type. For example, they realised the need to consider specification requirements while choosing the writing prompt scenario. They demonstrated the ability to see an item as a whole in a synergy of all its parts, and they saw connections between different specification requirements. Most importantly, they realised the importance of specifications in item-writing and reported paying attention to all aspects of them.

Post-training, *'high-achievers'* retained most features of their pre-training approach while also refining and improving on it. They kept paying attention to all aspects of specifications, while their understanding of the specification requirements deepened and became more nuanced. They developed some effective ways of working with specifications, for instance they used two screens and applied an iterative item-writing process by regularly checking their item drafts against specification requirements. They were able to provide a detailed account of their item-writing process, and they reported some effective item-writing approaches not observed prior to the training. For example, they introduced a preparatory stage to writing items. Their understanding of the inter-relation between different aspects of specifications, which was already considerable pre-training, improved and became more detailed. They reported using the knowledge and skills practised during the training, and they prepared for the post-training item-writing by revising relevant training material. They also critically reflected on the items they wrote pre-training.

Profile C included those participants who wrote good-quality items pre-training but who produced lower-quality items after it. These participants demonstrated most of the characteristics of *'high-achievers'* with two important differences. Firstly, some of them seemed to have placed a lot of emphasis on a particular specification requirement or several requirements. The emphasis must have occurred as a result of the training and, in a way, might be seen as an indication of learning. However, for novice item writers, such skewed attention resulted in reduced attention to other specification requirements and, consequently, lower scores on the relevant criteria. Secondly, some of these participants, feeling more confident after the training, adopted a more daring attitude to item-writing by exploring a wider range of topics, genres, and other aspects as described in the specifications. They wanted to produce something markedly different from the example item but, having little item-writing experience beyond induction training, they did not have enough ability to produce very original items. Their efforts resulted in some inappropriacies such as a wrong input message genre or the item being biased, which led to an overall lower item quality.

Some subjectively-scored requirements were particularly difficult for participants to deal with before attending the course. They were also the ones most discussed in the interviews. Among them were the construct for grammar items, distractors for grammar and listening items, and input characteristics for writing and listening items.

The construct requirements were generally managed well both for writing prompts (the ability to write a formal email of complaint) and listening tasks (the ability to locate and record specific information from a spoken monologue). However, this was not the case for grammar items. Only those participants who understood the grammar construct well and considered ways of realising it in items were successful. There was also a surprising mismatch between many participants' declared lack of grammar knowledge on one hand, and their perception of producing grammar items as easy. A2 grammar items in particular were considered easy to produce, which might point to a misconception, among novice item writers, about the simplicity of writing items to 'simple' grammar constructs. The perceived simplicity might have led to excessive confidence that the construct was clear, unwillingness to consult grammar reference material, as well as less time spent on actual item production, including choosing strong distractors. At the same time, the complexity of the C1 structure made participants research it and spend more time on the item, which resulted in higher scores. The knowledge of grammar and, in fact, any other aspect of language proficiency is normally taken for granted in item-writer recruitment, the assumption being that the applicants are linguists (in the broad sense). The findings, however, suggest that novice item writers might require some explicit grammar input.

One of the most difficult requirements for the participants in this study was the distractor requirement for grammar and listening items, which was often discussed in the interviews. The best distractors for the listening task, for example, were produced by those participants who considered the distractor requirement while selecting the text scenario and who incorporated distractors into the text during its writing and not *after* the text had been written. However, coming up with effective ways to comply with subjectively-scored requirements seems to have been more difficult than learning to comply with objectively-scored requirements. This suggests that most novice item writers might require explicit training on how to approach the production of specific item types, as only some novices seem capable of finding the right approach by themselves.

Input message (writing prompt) and input text (listening task) characteristics were among the most discussed topics by participants before and after the training. Participants who produced high-quality writing prompts seem to have realised the importance of choosing the right input message scenario from the start, because the suitability of the scenario seems to have had a decisive effect on the overall item quality. Scenario plausibility was the main concern for *'high-achievers'*; they also selected the input message scenario with all of the specifications in mind. After the training, they added a 'preparatory' stage to aid their item-writing process. For example, Olivia created a list of possible scenarios and checked each scenario's vocabulary against the word frequency requirements.

The participants' ability to create authentic-sounding listening texts greatly improved after the training and is one of the biggest observed successes of the item-writing training in this study[10]. The most effective item-writing technique proved to be recording a listening text from

[10] This study pre-/post-design is limited to the language skills and item types discussed in this chapter. The study did not analyse participant item-writing skill development on all language skills / item types taught during the training.

content points, or semi-scripting. Overall, it seems that creating authentic-like input for listening tests is an item-writing skill susceptible to fast improvement. However, it also seems that a few participants focussed excessively on the authenticity requirement, which brought about an unwanted drop in quality on some other related criteria.

4.4 Role of the training (RQ3)

This section reports findings related to the third research question, *"What role did the participants perceive the training played in their item-writing skill development?"* First, quantitative and qualitative findings from four feedback questionnaires administered to participants throughout the course are discussed in Section 4.4.1, grouped into four sub-sections: training materials (4.4.1.1), training activities (4.4.1.2), the course structure (4.4.1.3), and the use of technology to facilitate online learning (4.4.1.4). Second, qualitative findings from the post-training interviews in the part where participants gave feedback on the training are outlined in Section 4.4.2 which is organised into four sub-sections identical to the ones that present findings from the feedback questionnaires.

4.4.1 Findings from feedback questionnaires

As discussed in the Methodology Chapter (see Section 3.5.2.1), participants completed four anonymous feedback questionnaires: one questionnaire after every two modules of the training, and the Final questionnaire upon submission of the post-training assignment. The first three questionnaires each focussed on a specific aspect of the training: training materials (FQ1), training activities (FQ2), course structure and the use of technology (FQ3). The Final FQ revisited all four aspects of the training allowing for comparisons of the responses that participants provided during versus after the training. The four sub-sections that follow present findings for each aspect of the training in turn.

4.4.1.1 Training materials

Upon completion of Modules 1 (Introduction to item-writing) and 2 (Producing grammar items) of the training, participants were asked to complete FQ1 (*Appendix 10*) to share their impressions of the training materials, both in general terms and in relation to specific materials used in Modules 1 and 2. The data obtained was both quantitative (in response to enforced questions with Likert-scale options) and qualitative (in response to enforced follow-up questions asking participants to justify their response choices). The Final FQ completed after the training reused the Likert-scale questions from FQ1 to ask participants about their general impressions of training materials throughout the whole course. Unlike FQ1, the final FQ contained only one follow-up question which was not enforced. Nineteen responses were registered both to FQ1 and the Final FQ.

FQ1 asked participants to evaluate the training materials for their usefulness, interest, user-friendliness, and quality on a scale from 0 to 10, with 0 being 'absolutely useless' and 10 being 'extremely useful'. The mean values of responses ranged from 6.80 for user-friendliness to 7.68 for quality (*Table 4-40*). Overall, participants had a positive impression of the training materials in Modules 1-2, (see *Figure 4-13*), with individual participants differing in their evaluations: the lowest scores of 1 to 3 were never selected, a small number of participants selected a lower score of 4, 5 or 6, and a similarly small number of participants selected a high score of 9 or 10. The majority of responses clustered within the range of 7-8.

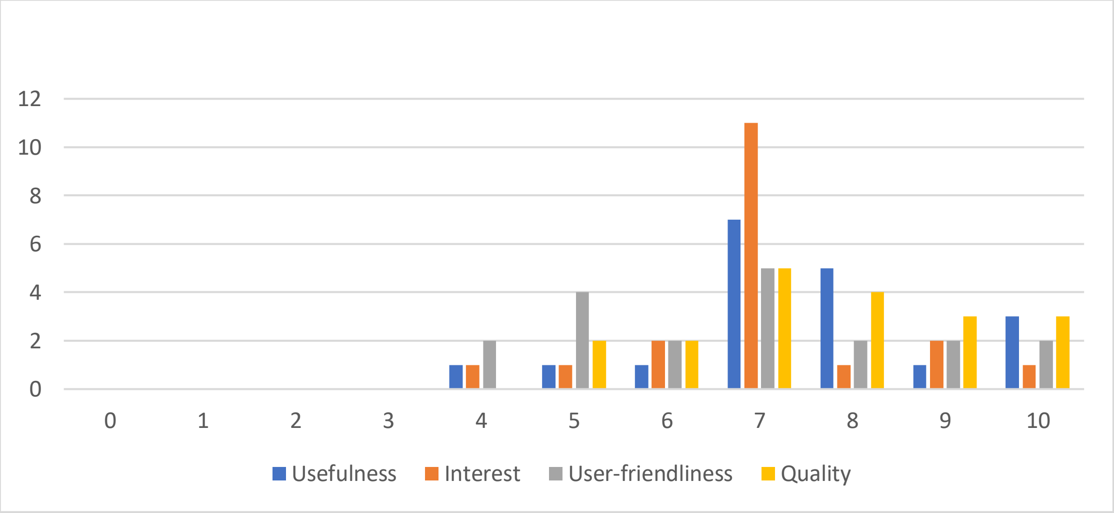


Figure 4-13. Participants' evaluations of the training materials in FQ1

Training materials' usefulness and quality received particularly high appraisal. Participants wrote that the materials *"fit the training aims well"*, provided clear explanations, and *"allowed for deeper understanding of the topic"*. The materials were generally perceived as well-produced, authoritative, and thorough. Many participants also described the materials as interesting and engaging, although more theoretical materials, such as academic papers, were perceived as somewhat dry albeit useful.

Materials' user-friendliness received somewhat lower evaluations (M=6.8) due to some problems related to the online mode of delivery. Most PowerPoint presentations (PPTs) used during the training were accompanied by a voiceover and some participants found that the voiceover and the slideshow were sometimes out of sync. Some also found that PPTs took time to download, which might have been because the participants were based in China: the restrictions on foreign internet traffic may have affected both the speed and the quality of access to the course. After the participants had reported their technical problems with PPTs', transcripts of all presentations were provided starting from Module 3. Some participants also found text documents difficult to open depending on the device they were using, which might point to the need for document provision in several different formats. Some participants' lower evaluations, as reported by the participants themselves, were a reflection of their lack of familiarity with online modes of training and of working with materials online:

> *For me, paper handouts would have been more effective … I know some course members printed materials. I don't have a printer. Probably if I had printed the materials, they would have worked better for me.*

FQ1 asked participants to evaluate specific materials used in Modules 1-2 on a five-point scale from 'totally useless' (=0) to 'extremely useful' (=5). PPTs were generally found to be useful, effective, practical, clear, applicable, and interesting, with mean values ranging from 4.05 to 4.42 for individual PPTs (average M=4.16). In comparison, Davidson and Fulcher's (2007) article which had been used to generate a discussion on the role of the CEFR in item-writing received mixed evaluations (M=3.31). Some participants thought it was informative and useful, a "*great introduction to important issues"* that *"familiarizes us with ongoing issues in assessment",* while many participants believed it was vague, heavy-going, long-winded, and rather academic. One participant wrote that the article was *"an interesting read, but I think something more succinct aimed at a less academic audience might be more appropriate".* Similarly, while three participants asked for more articles *"in a similar vein to Davidson and Fulcher's … as extra reading material*", the majority asked for ones that were more practical.

Quizzes also received mixed evaluations, with the mean scores ranging from 3.53 to 3.58 (average M=3.55). Some participants thought the quizzes were useful, practical, and provided immediate feedback, while others believed that the quiz named *Dos and Don'ts of Item-writing* was too vague with ambiguous answers and that the quiz used to check participants' familiarisation with the *Core Inventory* document was too easy.

Materials used to introduce practical activities, such as a worksheet with ten weak grammar MCQs for analysis (M=4.37) and a worksheet with grammar item specifications for item-writing practice (M=4.63) were perceived as the most useful: they were described as stimulating, challenging, very targeted, and helpful. One participant commented that *“this is exactly what I want from the course”.* When asked about other types of materials that might be helpful in developing their item-writing skills, participants mostly requested *“more examples of good and bad items”* (six participants) and *“more item-writing practice”* (five participants). Only three people, as discussed above, requested more theoretical input that would underpin item-writing practice. One participant also asked for grammar reference book recommendations because *“everyone only pretends to know their grammar.”*

It should be kept in mind that the above views were elicited after two Modules only. The Final FQ, administered after the training, asked participants the same questions as discussed above (but did not request justifications for the responses). Mean values for the post-training responses were all higher compared to the ones provided after Modules 1-2 (*Table 4-40*), indicating that the participants’ evaluations of the training materials were becoming higher as the course progressed. This finding is supported with a comparison of individual responses provided for FQ1 (*Figure 4-13*) versus Final FQ (*Figure 4-14*): after the training, score 4 was never selected, scores 5-6 were selected much less frequently, while score 8 was selected much more frequently.

Table 4-40. Participants' evaluations of training materials: Mean values

	Usefulness	**Interest**	**User-friendliness**	**Quality**
FQ1	7.53	7.05	6.80	7.68
Final FQ	8.21	7.79	7.26	7.84

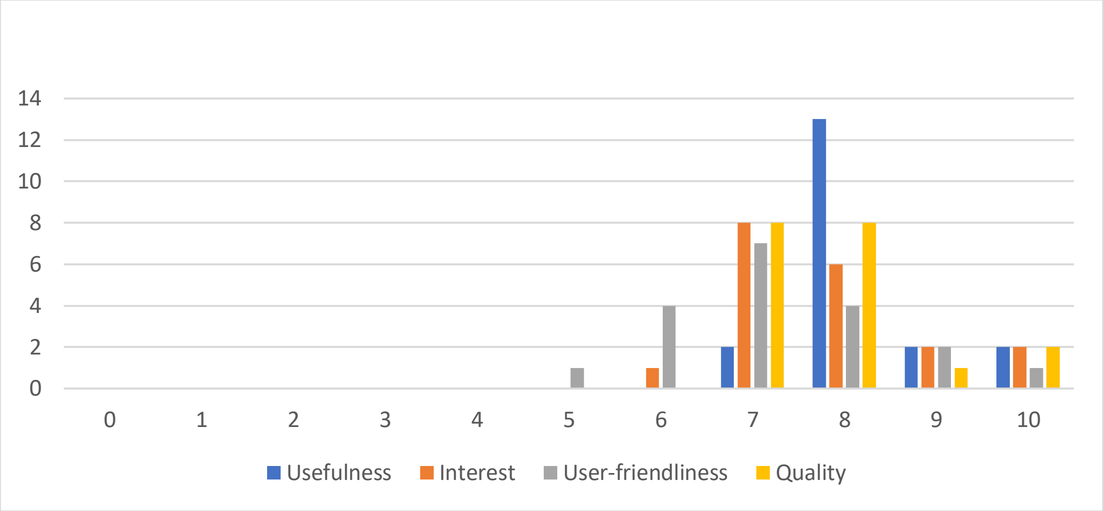


Figure 4-14. Participants' final evaluations of the training materials

The Final FQ (*Appendix 13*) contained one optional question on participants' suggestions for materials' improvement. Eight participants provided responses that largely repeated the suggestions expressed in FQ1: more *"weak and strong"* item examples, *"worksheets identifying errors in items with clear answers and explanations"*. One participant also asked for item-writing guidelines that would detail the item-writing process for different item types and language skills, such as *"taking a reading text from its raw found form to a good usable item, the processes involved in creating items".* Most suggestions, however, were related to the online delivery: hard copies of all materials in addition to digital copies, presentation software different from PowerPoint to seamlessly integrate the audio and slides, and a more comprehensive way of organising materials online.

4.4.1.2 Training activities

FQ2 (*Appendix 11*) was administered after Modules 3-4 and asked participants to provide a general evaluation of the training activities, as well as to reflect on the usefulness of specific activities in Modules 3 and 4. Similar to FQ1, the data obtained was both quantitative and qualitative. The Final FQ reused the Likert questions in FQ2 without asking for response justifications.

Participants evaluated training activities for their usefulness, interest, and user-friendliness on a scale from 0 to 10 (*Table 4-42*). Similar to what was found for training materials in FQ1, the activities' usefulness received the highest mean score (M=8.04), with most individual scores clustering in the range of 7-9 (*Figure 4-15*). Participants wrote that the activities were varied, provided a good combination of theoretical and practical tasks, gave *"lots of opportunity for* *practical item writing and discussion in groups"*, *"raised awareness of the skills needed"*, and *"showed very well the rationale as well as the practice of item writing".*

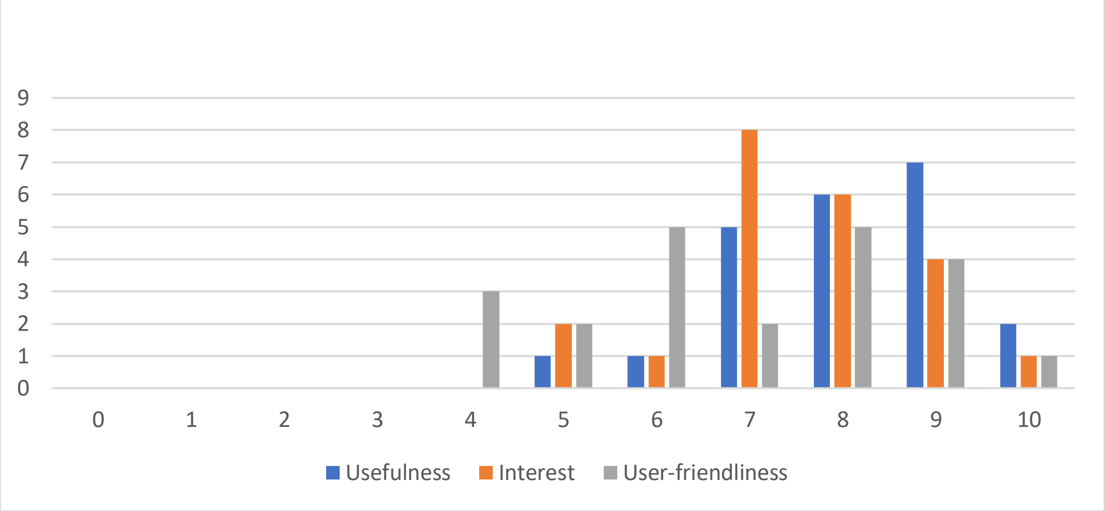


Figure 4-15. Participants' evaluations of the training activities in FQ2

Participants generally found the training activities interesting (M=7.54), and it seems that individual participants' interest largely depended on the type of item / language skill they preferred. Some found activities in Module 3, which focussed on vocabulary items, more interesting because they liked writing vocabulary items; other participants enjoyed Module 4 more because they preferred producing speaking and writing prompts.

User-friendliness of the training activities received somewhat lower evaluations (M=6.90) which, similar to what was found for FQ1, was linked to the use of technology. Participants had to download activity worksheets from *Edmodo*; they also used *Wechat* for group discussions and online tools (*Cohmetrix* and *Lextutor*) to produce items. If any of the technology failed to work, the activity affected was perceived as less user-friendly. The extent of the problem seemed to vary among individual participants depending on the participants' Internet connection and digital literacy – some *"had no difficulties"* using the technology, while others were struggling to download documents or make online tools work.

FQ2 also asked participants to evaluate specific activities used in Modules 3-4 on a five-point scale from 'totally useless' (=0) to 'extremely useful' (=5). Although the appraisals discussed above reveal that participants generally felt the combination of theory and practice was beneficial in their learning about item-writing, individual practical item-writing activities in Modules 3-4 received higher evaluations compared to the activities related to item-writing theory. Participants felt that writing a multiple-matching vocabulary task (M=4.62) and producing speaking and writing prompts (M=4.67) were the most useful activities in Modules

3-4. Participants believed these activities were challenging, rewarding, interesting, and generated useful feedback for future improvement. Practical activities in preparation for item-writing, such as adjusting the vocabulary difficulty of a text (M=4.54) or analysing weak items (4.19), were also evaluated highly by participants, who believed they were acquiring useful skills applicable not only to item writing but also to analysing and producing teaching materials.

Tutor feedback on items produced during the course received highly positive evaluations. Individual feedback (M=4.65) was perceived as indispensable. Participants particularly appreciated the fact that the feedback was detailed and thorough and followed the *Quality Review Checklist* format. The responses contained two tutor feedback-related suggestions: to provide *"more info about the good aspects of my items"* and to extend the practice-feedback loop by allowing participants to improve their items based on tutor feedback, resubmit, and then repeat the cycle until the item had no faults. However, participants also recognised this was not fully realistic given the time constraints. Group feedback (M=4.29) in the form of task summaries was also perceived as useful:

> *Seeing how others approached the tasks was interesting, and also made me realise some small things I could have done better; I really enjoy task summaries, it's a good way to conclude and pinpoint what's important. Also I like to go back to this now and then as a reminder…*

The attitude to peer-feedback (M=3.86) seemed to differ substantially among participants. Some evaluated it very highly:

> *The constructive peer feedback proved to be very useful. This is my favourite part of the course. I learnt a lot from other participants' tasks, mistakes and feedback they received.*

Others, however, wrote that peer-feedback had limited value because the discussion groups where participants reviewed each other's' items varied in their amount of activity:

> *This was very useful at times, but it depended on how vocal other group members were. Sometimes I received no feedback; at other times we had good discussions.*

The activity related to the theory of item-writing, whereby participants had to read and discuss a chapter about lexical competence (Meara, 1996) received somewhat lower evaluations (M=3.45). Similar to what was found for FQ1, the opinions were divided: some participants enjoyed reading and then discussing the chapter because it was thought-provoking, relevant to the training aims, and provided them with good ideas. Discussing the chapter in groups was perceived by these participants as useful because it allowed them to *"hear other's views on the text"* and get their *"understanding on the text confirmed by others"*. However, other participants found the reading boring, long, and time-consuming because they were *"not really into theory"*.

FQ2 asked participants to express a preference for the mode of interaction in doing training activities, choosing among working individually, in groups, or as a combination of both. The results (*Table 4-41*) reveal a clear preference for the combination of individual and group work because it is good for learning, provides variety, and *"reflects two key stages of item writing, the creation (which is usually done individually) and the review (which usually involves at least interaction between the reviewer and the writer)".* It also seems that participants' preferences depended on their personal disposition: some said they generally preferred "*working alone*", while others needed group collaboration to stay motivated. Some participants also commented that the quality of group work depended on group members and suggested that groups should be *"moderated more actively"* to stimulate participation, for instance a group leader should be appointed, or the tutors should be more active in encouraging individual participants to post. Some participants also linked their lack of participation in group activities to problems with technology and/or busy work schedules.

Table 4-41. Participants' preferences for the mode of interaction

	Most preferred	**2nd preferred**	**Least preferred**
Working individually	7	7	7
Working in groups	1	7	13
A combination of both	13	7	1

The final two questions in FQ2 asked participants for suggestions on how the training activities could be improved and what other training activities could be used to help the participants develop their item-writing skills. Three participants said they were *"very happy with everything"* as it was. Suggestions for improvement included: using an alternative technology that would be more accessible from China, more activities analysing good and weak items, more feedback on items, and keeping participants in the same groups throughout the course because *"it may be more effective for members to develop deeper relationships".* The suggestions for other types of activity included: face-to-face training, presentations whereby the tutors would *“talk through their mental process of creating an item”*, and peer-review in pairs:

> *…with one writer and one reviewer role; I think you would get more meaningful feedback from the reviewer in this case because, one, the reviewer would have proportional more time with only one item writer's work to review, and, two, there would be clear allocation of roles and responsibilities.*

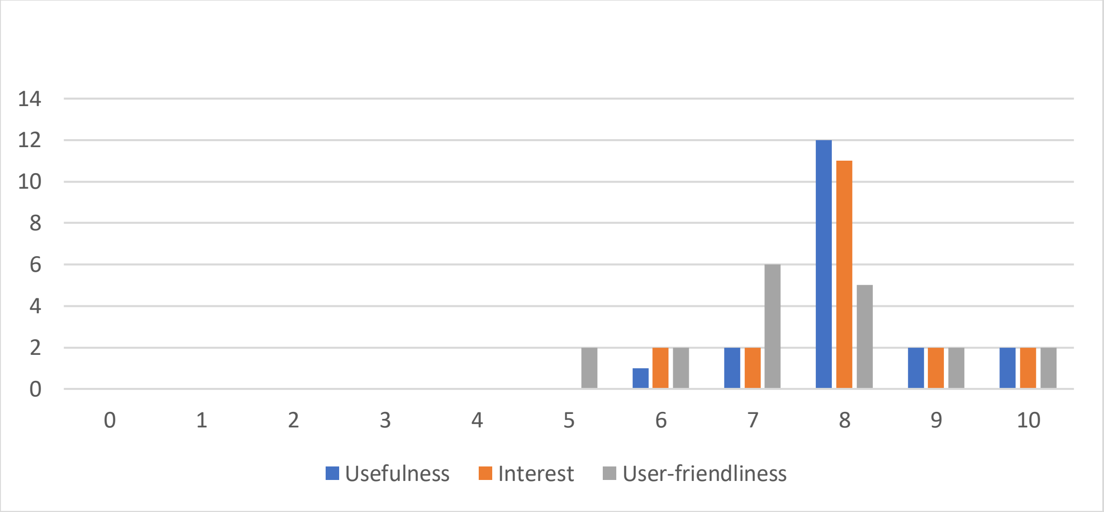


Figure 4-16. Participants' final evaluations of the training activities

Table 4-42. Participants' evaluations of training activities: Mean values

	Usefulness	**Interest**	**User-friendliness**
FQ2	8.04	7.54	6.90
Final FQ	8.10	8.00	7.47

The Final FQ saw still higher appraisal of course activities’ usefulness, interest and user-friendliness (*Table 4-42*), with no score 4 and much fewer score 5 awarded (*Figure 4-16*). The optional question asking for suggestions on improvements to the activities attracted six responses: one participant said the course was fine as it was, four participants offered suggestions regarding group discussions, and one participant asked for optional webinars to complement the asynchronous mode of study. The suggestions on group discussions mostly repeated what was already reported above: the participants felt groups had to be monitored more closely *“to force people to make the time to post when they are tired”*, while the discussions themselves had to be *“more structured”.*

4.4.1.3 Course structure

FQ3 (*Appendix 12*), which was administered after the last Module of the course and completed by 21 participants, aimed to evaluate the course structure according to several parameters: clarity, flexibility, and pace (evenness and speed) on a scale from 0 to 10. Participants generally thought the course was well-structured (M=8.24, *Figure 4-17*) because (1) it followed a logical progression from theory to practice, and from simple to complex; (2) it built on the skills acquired in earlier Modules, *"so knowledge, conventions and skills acquired could be re-used"*; (3) it offered a balanced combination of *"self-study, group discussions, sharing written items and peer reviews";* (4) each Module was organised in a similar way *"starting off from lead - in a form of narrated ppt, followed by independent then group tasks",* with all tasks being well-linked. Participants also highly valued the fact that reflection on the pre-training assignment was incorporated into course work:

> *I thought that the idea of learning about an aspect of item-writing, followed by self-reflection on the pre-course item, then combined with the actual feedback of our pre-course item and finishing off on improving the same item (using recently gained knowledge) was excellent! I will definitely use this set up in the future if I am involved in a course design.*

Three suggestions for improvement were given: leaving *"the academic papers till later in the course"* because they intimidated some participants in the earlier Modules, having smaller *"bite-sized"* units *"to help us get into the routine of thinking about it* [the item-writing] *every day"*, and providing some input *"about the item-writing market and job opportunities".*

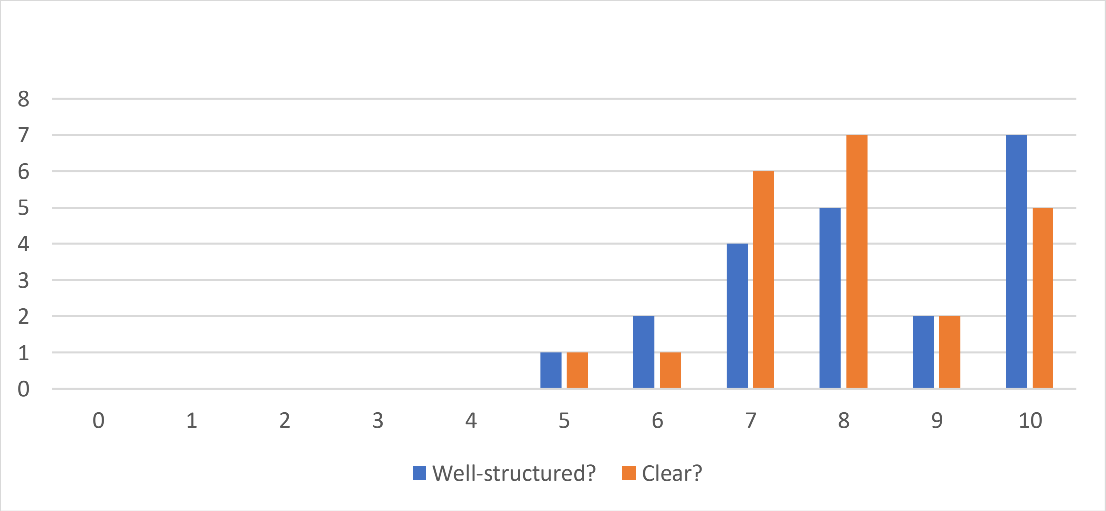


Figure 4-17. Participants' evaluations of the course structure in FQ3

The course structure was perceived as clear by the majority of participants (M=8.14) because each Module followed a similar structure*,* participants *"were given module summaries beforehand, a timetable"*, so they *"generally knew what we were doing each week and why".* The very few negative comments concerned the online learning platform (*Edmodo*) because of the way it displayed the materials.

Ninety percent of respondents found the course flexibility and pace appropriate. However, the perception of what was appropriate differed among individual participants. For instance, 28% thought the course was *'appropriately fast'*, 5% - *'appropriately slow'*, while 57% thought the course was *'neither fast nor slow, which was appropriate'* (*Table 4-43*). In terms of flexibility, participants appreciated the fact that the tutors were flexible with deadlines and took individual circumstances into account. Participants also found that, although the course was demanding, they could generally fit it into their work schedule. The alternations of *"hard and easy weeks"* gave participants *"natural breaks in intensity",* and participants appreciated the fact that more time was allowed to write items for receptive skills because these were unanimously regarded as more difficult and time-consuming to produce. Among the suggestions for improvement were setting deadlines twice a month instead of every week and giving participants a *"heads-up"* that the last two Modules of the course, dealing with listening and reading items, would require *"a much larger work commitment".*

Table 4-43. Participants' evaluations of the course flexibility and pace

	FQ3	**Final FQ**
Course flexibility		
• Appropriately flexible	76%	84%
• Not flexible enough	10%	5%
• Too flexible	0%	0%
• Appropriately inflexible	14%	11%
Course pace (evenness)		
• Appropriately even	81%	79%
• Appropriately uneven / varied	9.5%	16%
• Too even	0%	0%
• Too uneven	9.5%	5%
Course pace (speed)		
• Appropriately fast	28%	11%
• Appropriately slow	5%	5%
• Neither fast nor slow, which was appropriate	57%	68%
• Too fast	10%	5%
• Too slow	0%	11%

FQ3 also prompted participants to evaluate the usefulness of Module 6's structure on a scale from 0 to 5, in particular its task sequencing, use of interactive activities, and flexibility and pace. The task sequencing (M=4.33), flexibility (M=4.14) and pace (M=4.05) were positively evaluated, which aligns with the participants' perceptions of the course structure in general. The mean value for the use of interactive activities was somewhat lower at 3.38, which also reflects the findings about group activities in FQ2. Although many participants felt it was *"very useful to try and evaluate other people's items"* and *"the contact with others was extremely useful and productive",* respondents' perceptions were affected by individual experiences, which varied a lot. Those participants whose groups were less active or who did not receive peer-feedback on their items, evaluated the interactive activities as less useful.

The Final FQ responses (*Table 4-44*) confirmed participants' satisfaction with the course structure. The course flexibility and pace were particularly highly evaluated (*Table 4-43*) with 95% of participants finding them appropriate.

Table 4-44. Participants' evaluations of the course structure: Mean values

	Well-structured?	**Clear?**
FQ3	8.24	8.14
Final FQ	8.10	8.00

Only five participants offered suggestions for course structure improvements in the Final FQ. All of them wanted the course to *"run longer"* with more breaks in-between the Modules to allow for catch-up, and more fluid deadlines. A longer course would also allow for *"a second submission after the first QR review to really deepen the learning and have more feedback".*

4.4.1.4 Use of technology

As well as collecting responses about the course structure, FQ3 also invited participants to evaluate, on a scale from 0 to 10, the use of technology on the course in terms of its usefulness (M=6.80), supportiveness (M=6.95), and user-friendliness (M=7.05). Overall, the use of technology received slightly lower evaluations compared to other aspects of the course (*Figure 4-18)*, with participants focussing on individual pieces of technology in their responses, rather than providing an overall evaluation.

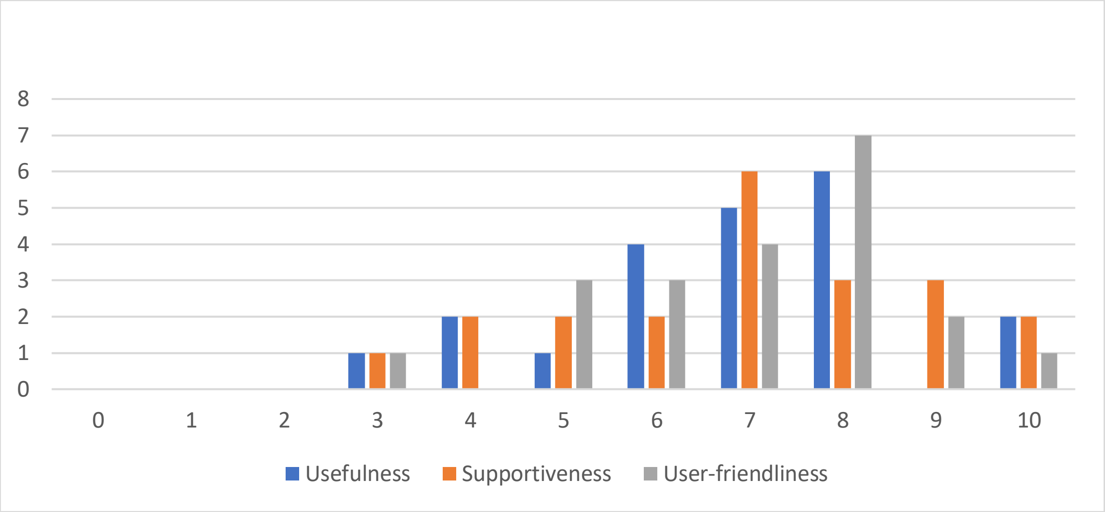


Figure 4-18. Participants' evaluations of the use of technology in FQ3

A range of technology was employed during the course: *Edmodo* as a Learning Management System (LMS), *Wechat* as a platform for group discussions and peer-reviews of items, email to submit the final version of the items to the tutors and to receive individual feedback, *Cohmetrix* and *Lextutor* to check items for readability and vocabulary frequency, and PowerPoint presentations for tutor input. Participants had a chance to comment on individual uses of technology in the second part of FQ3, which asked participants to evaluate, on a scale from 0 to 5, how technology was used in Module 6 of the course.

Edmodo

Three respondents commented positively about *Edmodo* because it was simple, easy to use, could be accessed both from a PC and a phone, and was *"fine for a free platform"*. The majority, however, held more negative views, explaining that *Edmodo* was slow in China and required a VPN, is *"unsophisticated", "underwhelming",* and *"child-focussed"*. The library, an *Edmodo* feature which allows for the storage of all course materials in one place, was perceived as not user-friendly because it *"seems to put the files in a silly random order".* This probably explains why the use of *Edmodo* as a library to store Module 6 materials received the lowest evaluation (M=3.28). Participants wrote that "*any FTP server would do*", such as Dropbox or Google docs. The use of *Edmodo* to introduce module aims and activities (M=3.57) and to access task summaries (M=3.62) was evaluated more positively because it was easy and straightforward to access the information. However, some participants said they would have preferred to have module introductions and summaries emailed to them.

Wechat

Opinions on *Wechat* were almost equally split: half of the responses praised *Wechat* as a good platform for group activities, appropriate in the Chinese context and working very well, *“fine as a forum for reviewing and discussing each other’s work”.* Other participants, however, thought that *Wechat* was less suitable for group discussions, with various reasons provided: *Wechat* is mostly used on tablets and mobile phones and, although it’s possible to post to *Wechat* from a PC, it is not convenient; *Wechat* is not convenient for reading long posts; shared files expire after several days; and *“messages received on one device aren't visible on any other devices”.*

The use of *Wechat* in Module 6 for peer-feedback (M=3.90) was positively evaluated by many participants who said that “*Wechat shines for this purpose”* and was *“a better choice, as comments were instant and sometimes could involve discussion”.* As a medium for the text-mapping activity, however, *Wechat* was evaluated somewhat lower (M=3.71) because it is “not *useful for activities where everyone is expected to come up with more or less the same”* and *“…can be messy when catching up with discussions and trying to read people's posts”.*

Email

The use of email received highly positive evaluations, both for the course as a whole and in Module 6 to submit listening items produced by participants (M=4.28) and to receive individual feedback from the tutors (M=4.19). Email, as a familiar medium, did not pose any problem to participants and was thought to be *“normal”* and *“appropriate.”* However, several participants suggested eliminating email communication to reduce the amount of different technologies used on the course. Some thought that *Wechat* could be used instead: *“As most activities are done over Wechat it might make more sense to send completed tasks to you via Wechat as well”.* Others would have preferred participant-tutor communication to happen within the LMS: *“It would be good if we can email them* [the tutors] *within the same platform where we get our materials so everything is stored in one place”.*

Item-writing tools: *Cohmetrix* and *Lextutor*

Most participants complained that they could not access *Cohmetrix* from China and requested a different readability tool. As for *Lextutor*, it was described as *“very poor”.*

PowerPoint presentations

Most participants, when commenting on materials for FQ1 (see *Appendix 10*), described PPTs as a useful and adequate way of providing course input. Comments in FQ3 revealed some technical difficulties that several participants had experienced using PPTs: some participants found it difficult to download narrated PPT files, the voiceover was sometimes *"out of sync,"* and one respondent wrote that "*PowerPoint would jump to the next slide before the audio was finished, leaving me to have to go back and watch every slide twice"*. Also, when participants tried to use a built-in PowerPoint player in *Edmodo*, it could not run narrated PPTs.

FQ3 also asked participants to provide suggestions on how the use of technology could be improved to help develop their item-writing skills, and what other technology could be used for that purpose. Three participants suggested using one platform for all course-related activities. Several other participants recommended replacing *Edmodo* with *"a more professional-looking"* LMS, such as *Moodle*, or with a shared folder. Several participants suggested using *"a forum format for discussions"* hosted on an LMS such as *Moodle*. Most others, however, thought that *Wechat* was appropriate for the purpose. One participant suggested using Google docs to write items as a group because *"a co-written task (reading or listening) could be fun to put together".* A number of participants also suggested including an opportunity for synchronous communication: optional webinars using WebEx or Zoom software, live Q&A sessions with course tutors who would *"offer some time slot to discuss any questions we have about each module",* as well as using video-conferencing technology *"to facilitate some pair-work item-writing".* One respondent furthermore suggested using videos recorded by course tutors: "*This video introduces a particular module and shares the course leaders' personal experience of being an item writer".*

The Final FQ resulted in similar evaluations for the use of technology on the course (*Table 4-45; Figure 4-19)*, with ten participants providing suggestions on how the use of technology could be improved. All suggestions were identical to the ones discussed above.

Table 4-45. Participants' evaluations of the use of technology: Mean values

	Usefulness	**Supportiveness**	**User-friendliness**
FQ3	6.80	6.95	7.05
Final FQ	6.79	6.68	6.53

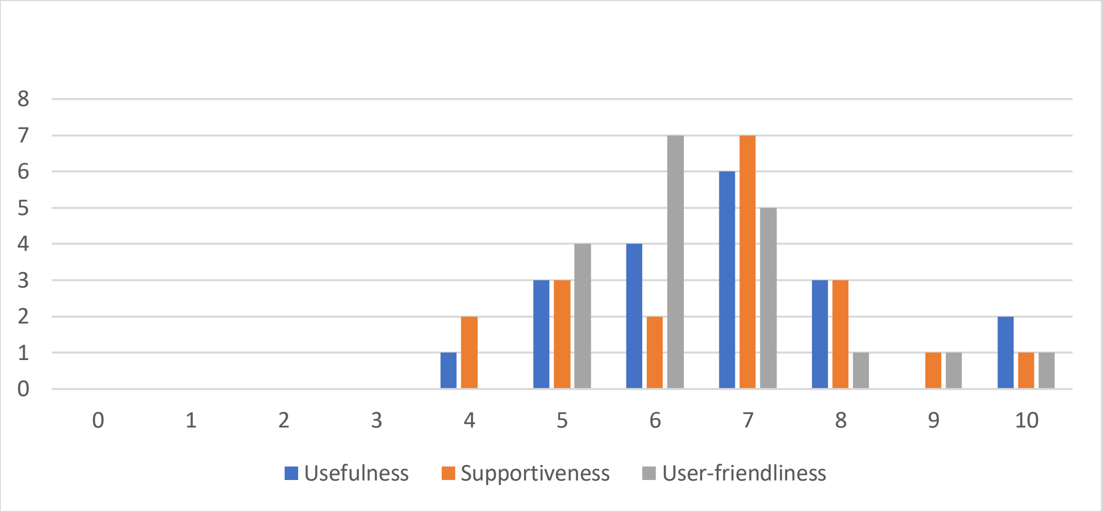


Figure 4-19. Participants' final evaluations of the use of technology

The last question in both FQ3 and the Final FQ asked respondents to share any other thoughts about the course. All responses contained high praise for the course and thanks to the course tutors. Some examples are:

> *I was very impressed. When it comes to course content this is top-notch! Beside the technology I have no complaints.*
>
> *The course has been very interesting, and I have learned a lot, as well as revisiting some professional areas that I haven't touched upon since I was an MA TESOL student. Thank you very much for all your hard work.*
>
> *This was one of the most practical courses I have ever completed, and I enjoyed writing the items despite not always being good at it. I would be happy to continue this course for much longer and keep writing the items.*

4.4.2 Findings from post-training interviews

As reported in the Methodology Chapter (see Section 3.5.2.1), interviews conducted after the training included a discussion of the course helpfulness (or lack thereof) in developing participants' item-writing skills. Without mentioning any specific course materials or activities, the interviewer asked participants to reflect on what aspects of the training might have helped them in writing items for the post-training assignment, as well as how the training could be improved to better help the participants in writing items. The findings are presented in four sub-sections: participants' reflections on the training materials (4.4.2.1), training activities

(4.4.2.2), the course structure (4.4.2.3), and the use of technology (4.4.2.4). The sub-sections are identical to the ones used to present findings from the feedback questionnaires, which enables comparison of the two types of data. The identical structure is not deliberate but resulted from thematic coding of the interview data, with participants' comments neatly following the themes of the four feedback questionnaires (see *Appendix 9* for the list of codes). A potential explanation is that the participants sub-consciously followed the evaluation categories they had previously encountered while responding to the feedback questionnaires.

4.4.2.1 Training materials

Participants' overall impression of the training materials was favourable: the materials were characterised as useful and *"applicable to the task of item writing"* (Arthur). Two presentation issues were raised: Liz said it was *"very hard to keep track of all the different files, so while things were useful it was hard to go back and find them again",* while Henry suggested that *"it would be nice putting it* [all course materials] *into a handbook of sorts so that you can refer to it".*

Thirteen participants commented on the input materials that discussed theoretical foundations of item-writing, including PPTs, documents such as the CEFR and the *Core Inventory*, and academic articles. All 13 participants found this input useful, for example:

> *…there's also a rationale behind those resources, not just 'this is what we use'… definitely for awareness raising it was great* (Emily)
>
> *…the rationale behind things is, of course, extremely useful, and the way the rationale is explained obviously makes it a lot clearer to see what is the process of the writing* (Jake)

Participants particularly praised input on the construct underlying language tests and characteristics of different item types. For example, Olivia said the information was *"very practical and useful and it kind of felt like 'OK, I can do something with this' – or at least I could try to incorporate a piece of information into my practice right now".*

Three participants said that the two papers - Davidson and Fulcher (2007), and Meara (1996) - they read during the course were helpful. For example, James said they were *"great, very interesting… this was all new to me".* Nathan suggested having *"an additional reading list"* for future item-writing courses because *"there's one or two keen readers there, keenly attentive people that might be interested".* It should be noted that, similar to what was found for the feedback questionnaires, only a limited number of participants enjoyed reading the papers, while some others found such reading boring: *"Theory is good and important, but it's more fun doing it"* (Henry).

Most participants who talked about the training materials found input on the CEFR, including the *Core Inventory* document, very useful because *"before there was no real sense of grading language"* (Ted). The input helped participants better understand the proficiency levels and, consequently, target the items they produce at the right level.

Eight participants commented on the PPTs saying they were useful, *"very specific, practical, good balance between bullet point theory"* and practical item-writing (Mason). Participants appreciated that, besides theoretical input, the presentations contained a lot of item examples, for example Olivia recalled that the PPTs *"gave examples of items that did and did not work and why, that was very useful".* Several participants said that even more examples of good and weak items would be welcome because *"you can never have enough examples"* (Mason). Logan also mentioned that studying the presentations helped him navigate through each new Module of the course:

> *The PowerPoints were good because I think they allowed us to focus on what we were actually supposed to be doing … I'd begin to panic thinking there's a lot to do, but just going to that first PowerPoint then made sense, I'd read it, listen to you and then start the task, and it all sort of fell into place after that.*

Participants' opinions on how the PPTs would ideally be produced differed. Arthur would have liked the input presented in a document rather than a PPT format because he liked *"reading more than listening to PowerPoint presentations"*. Josh suggested having less voice-over and more visuals *"like a picture or a diagram"*. Logan, on the contrary, said that *"having a voice out there is better than reading it…".*

Only Jake mentioned quizzes, saying *"things like the quizzes … for me not quite so useful I have to say*", which reflects the somewhat lower appraisal the quizzes received in FQ1.

4.4.2.2 Training activities

In the interviews, the participants discussed group activities (including the ones done as preparation for the item-writing practice), item-writing practice itself, and peer-/tutor-feedback.

Generally, participants found all group activities useful, for example Lucy said, *"I find it useful sharing different ideas … and seeing how people approach the task".* Adam mentioned the advantage of having more than one person to think over issues, because *"everyone … they just* *miss something that's obvious like one of the key points in the PPT that doesn't cross their mind when they write their homework"* so they need a group member to point this out. Participants were allocated to a different group for each Module to allow them to communicate with everybody else on the course and to avoid being placed in an inactive group throughout the whole course. However, Mason felt that changing groups each Module was not beneficial, and he gave a different suggestion:

> *What I found is that there's a few people in the group who I kind of got on well with and we were sort of clicking really well as a group and there were others that were less so … when you move groups around … first of all there's a kind of familiarization period … a lack of confidence or worrying about upsetting people so I tend to pull punches … so the quality of peer feedback was probably not as good as it could have been... So maybe I think my suggestion would be DO change, it's important that you change the groups, give people that opportunity, but maybe don't change every module.*

Three participants discussed the item analysis activities whereby they were provided with weak items and, as a group, analysed them and provided suggestions for improvement. All three found such activities extremely useful because *"that compounded the idea of how to write items well"* (Daniel).

Ten interviewees commented on the item-writing practice saying it was *"clearly important"* (Emily), *"really-really useful"* (Joe), *"the essential part of the course"* (Josh), and *"the most useful thing we've done"* (Ted). One explanation for this praise was that the theoretical principles and item-writing rules *"become self-evident"* (Henry) while writing items. Nathan viewed *"just reading about it* [item-writing]" as insufficient because *"you need to actually do it and get feedback about what I'm doing wrong"*.

Ten and nine participants spoke about peer- and tutor-feedback, respectively. They generally regarded peer-feedback as useful and valuable*.* One explanation provided by Arthur was that

> *…working on an item and having the same specs as other people and comparing different ways of approaching the same kind of specifications is really helpful because you can see how other people approach it in ways that you had never thought about.*

Logan also speculated that working in teams is a normal item-writing process in operational language testing: *"I imagine people sitting in a room writing items asking each other what they think about it and then adjusting it and then finishing the task".* At the same time, Olivia found that *"… peer feedback was useful to varying degrees".* Participants also noted two potential problems with peer-feedback: not everybody in the group was equally active and forthcoming (Arthur), and they were reluctant to give negative feedback, particularly to unfamiliar participants in a new group (James and Mason).

All interviewees who spoke about tutor feedback found it very useful. They said the individual feedback on their items was *"very detailed"* (Lucas) and *"thorough"* (Nathan). Participants received tutor-feedback after each item-writing event, something which Logan found very helpful compared to the online courses he had done before where *"they expect everyone just to do their own thing and you get feedback way at the end,"* which he found less useful. Feedback timeliness was important to Logan because it allowed him to *"absorb"* the feedback before he attempted writing items again. Lucas expressed a similar opinion: *"…you make the wrong choices* [and] *you have to know what you did wrong so then the next time I do it, I know what to pay attention to".* Nathan also discussed tutor-feedback to the group as a whole, which took the form of a summary posted on the course platform at the end of each Module. Nathan said that, from this feedback, he was trying *"to absorb and digest … general errors that people were making .. as much as [he] could".*

4.4.2.3 Course structure

The course structure was commented on by six participants. All six provided highly positive evaluations saying the course was *"structured"* (Austin), *"well-designed"* (James) and *"very methodical"* (Josh). Some explanations included:

> *…there was a nice move from something quite simple like planning a grammar multiple-choice item to the more complex things at the end* (Emily)
>
> *…everything is linked together as well and so it's always continuous* (Logan)
>
> *I liked the balance between the individual and the group activities* (Mason)

4.4.2.4 Use of technology

Use of technology was the least mentioned topic with only four comments about the use of *Wechat* and *Lextutor*. Two interviewees thought that *Wechat* was a suitable platform to host group discussions. Daniel pointed out that all participants, while doing the course, were working full-time in a challenging job that involved a lot of travelling so *"it would be very tricky to say specific times that would work for everyone as part of the ongoing communication"* thus making synchronous discussions impossible. Additionally, Lucas noted that *Wechat* discussions worked better for the activities that required *"original and unique"* contribution such as peer-feedback or weak item discussions, whereas posting thoughts on an article was not as beneficial because *"one person wrote a lot of things at the beginning so there wasn't that much else to say"* – something that was also mentioned in FQ3. The two interviewees who discussed *Lextutor* said they were satisfied with the fact they had learnt to use this tool for item-writing purposes.

4.4.3 Summary

In the feedback questionnaires and post-training interviews, participants provided their opinions on the course materials, activities, structure, and use of technology. Training materials were generally very positively evaluated, with the evaluations being even higher after the course (Final FQ) than just upon completion of Module 2 (FQ1). The materials' usefulness and quality attracted particularly high scores. The slightly lower scores for user-friendliness can be explained by the online mode of delivery: some participants were not familiar with studying online, while restrictions on foreign Internet traffic in China resulted in slower access to some online item-writing tools.

Among individual materials, participants found the PPTs and worksheets with weak items for analysis particularly useful. The inclusion of materials on theoretical aspects of item-writing was appreciated by a small number of participants, who also asked for an extended 'reading list'. Others, however, felt the papers were somewhat heavy-going and requested more "succinct" input that would be "less academic". In terms of other material that they would have liked to use during the course, the majority of participants named (1) more item examples - both good and weak - with detailed explanations, and (2) an outline of the item-writing process for each individual item type.

Training activities were also evaluated highly positively, with the evaluations being even more positive after the end of the course. Participants particularly praised the balance of theory and practice, the variety of activities, and the large amount of item-writing practice. Although participants found the combination of theory and practice beneficial, they gave preference to practical item-writing activities which received the highest praise, followed by tutor feedback. Individual tutor feedback was thought to be more useful than group feedback, and peer-feedback received slightly lower evaluations with two problems reported: not all groups were equally active, and some participants felt uncomfortable about giving negative feedback on others' items.

A combination of group- and individual work was preferred to working only in groups or only individually. Most suggestions for further improvement concerned group work: participants would have appreciated fewer group rotations and closer moderation of the groups to ensure active and equal participation. However, individual participants' responses varied greatly depending on their preferences regarding types of activity, types of item to write, or mode of work.

The course structure received consistently positive evaluations throughout the course and in the post-course interviews. The course was considered well-designed, well-paced, flexible, with a clear progression from theory to practice and from simple to complex. Participants also appreciated the balance between different activity types and modes of interaction. There were few suggestions for improvements to the course's structure; these asked for the course to run longer with more breaks between Modules and more flexible deadlines.

The use of technology was evaluated somewhat lower compared to other aspects of the course. In particular, participants thought *Edmodo* was unsuitable as an LMS and wanted it replaced either with a file-sharing server such as Dropbox, or a different LMS such as *Moodle*. Participants also thought that the amount of different technologies used on the course should be reduced. One suggestion was to eliminate the use of email by conducting all communication via *Wechat*. Another was to move all course activity to a suitable LMS. Participants' opinions about the use of *Wechat* were split. Half of them thought *Wechat* was a suitable platform for group activities and discussions, while others would have preferred group activities to be carried out in a *Moodle*-style environment. *Wechat* was also viewed as more suitable for activities that required 'unique answers' such as peer-review, while discussing item-writing theory on *Wechat* was deemed less appropriate. Some suggestions included using synchronous communication in the form of optional webinars, Q&A sessions or group item-writing sessions.

It should be emphasized that the majority of the issues reported by participants were associated with the online course delivery and not its content or pedagogical approach. Participants themselves, however, must have seen the delivery problems as minor because they provided highly positive evaluations of the course as a whole in the Final FQ. They unanimously praised the course, the course tutors, and said it was one of the most comprehensive and useful courses they had ever attended.

Chapter 5 Discussion

5.1 Introduction

This chapter contains a synthesis of the key findings and discusses these with reference to the two learning theories introduced in Chapter 2 – the ACT-R cognitive learning theory and the CoP social learning theory. An integrated summary of the key findings is presented in Section 5.2. Section 5.3 discusses item-writing skills and their acquisition from the cognitive perspective, including the nature of item-writing skills (5.3.1), the process of item-writing skill acquisition (5.3.2), and item-writing skill acquisition for different item types and proficiency levels (5.3.3). The section ends with a discussion of the role of induction training in acquiring item-writing skills (5.3.4). Section 5.4 then moves on to discuss item writing as a social activity, arguing that item writers should be viewed as a community of practice (5.4.1) while item-writing induction training, in order to be effective, should strive to incorporate some features of 'legitimate peripheral participation' (Lave & Wenger, 1991) of novices in the item-writer community (5.4.2). The opportunities for 'legitimate peripheral participation' offered to participants in this study are then deliberated (5.4.3). Finally, Section 5.5 discusses this study's methodological contributions to the empirical research into item-writing training.

5.2 Summary of the main findings

Statistical analyses of item evaluations demonstrated that pre-training participants already had some ability to conform to item specification requirements, while post-training many participants produced better quality items. The Wilcoxon signed-rank test results showed that the total post-training scores for A2 and C1 grammar items, as well as B1 listening tasks, were statistically significantly higher than the corresponding pre-training ones. The improvement in the overall quality, however, was in a large part due to the statistically significant improvement in quality on the objectively-scored criteria, while the changes in the total scores on the subjectively-scored criteria were not statistically significant. Additionally, no significant differences in scores were detected for B2 writing prompts.

Gain ratio analysis revealed that score gains for the objectively-scored criteria were more uniform across the trainee cohort, compared to the subjectively-scored ones. With regard to the latter, four participant profiles were identified: (a) those whose pre-training items scored the lowest ('outliers') but whose post-training items were of much higher quality; (b) 'high-achievers' who produced good quality items before the training and whose post-training items were of even higher quality; (c) those whose pre-training items were of reasonably good quality, but whose post-training items scored somewhat lower; (d) those whose item quality improved following the training, but the improvement was not as drastic as for the 'outliers' or the post-training scores were not as high as for the 'high-achievers'. The analysis also revealed that participants did not demonstrate a uniform item-writing ability across all item types – one and the same participant could be an 'outlier' for the grammar items, a 'high-achiever' for the writing prompt, and scored lower post-training on his listening task, for example. Three of the above profiles (A-C) were then investigated in more detail using qualitative interview data, with main findings for each profile presented and discussed in Sections 5.3.2.1 -5.3.2.3 of this chapter.

Qualitative analyses of the pre- and post-training participant interviews revealed that participants generally found item-writing difficult, with listening tasks perceived as the most difficult and writing prompts as the least difficult to produce. Prior to the training, all participants largely relied on example items which they perceived as models to follow, while after the training most participants' attention shifted to studying the specifications. Most participants did not find complying with objective requirements difficult, although those who produced 'outlier' items pre-training seemed to need more guidance in interpreting the specifications and/or using item-writing tools. Despite a generally uniform increase in scores on objectively-scored criteria, there were several criteria that had lower mean scores following the training. Further analysis indicated that those criteria were competing with one or more subjective requirements; for instance, the word-limit for writing prompt input messages was competing with the requirement for the messages to be plausible and to sound authentic. It seems that pre-training participants focussed on complying with objective requirements while post-training their attention shifted to subjective requirements with the related objective requirement being overlooked. This was particularly true for participants whose items scored lower following the training.

Analyses also revealed that subjective requirements differed in their difficulty, with the grammar construct requirement, distractors for grammar and listening items, and input characteristics for writing and listening items being particularly challenging to comply with. Difficulties with grammar items' construct might be related to participants' grammar knowledge, while success at producing strong distractors seemed to depend on participants' use of effective item-writing approaches. For example, the best distractors for the listening task were produced by those participants who considered the distractor requirement while selecting the text scenario and who incorporated distractors into the text during its writing and not *after* the text had been written. However, not all participants were able to arrive at such effective approaches by themselves with many seemingly needing explicit training. The participants' ability to create authentic-sounding listening texts greatly improved after the training and is one of the biggest observed successes of the item-writing training in this study.

In their feedback, participants praised the balance of theory and practice, the variety of activities, and the large amount of item-writing practice they received during the course. Participants reported that input in language assessment principles helped them in clarifying specification requirements, while CEFR-related input was useful in producing items at different levels of proficiency. Most participants also noted that they preferred more 'succinct' input such as *PowerPoint* presentations over academic-style reading. Although participants found the combination of theory and practice beneficial, practical item-writing activities received the highest praise, followed by tutor feedback. Participants' most preferred mode of learning was a combination of individual and group activities. Many participants found group discussions and peer-feedback very useful, although individual perceptions depended on how active each group was and on individual participants' preferences.

The course structure received consistently positive evaluations with the course praised for being well-designed, flexible, well-paced, and having a clear progression from simple to complex and from theory to practice. The use of technology was evaluated at a slightly lower level, which seemed to be dependent on participants' access to a reliable internet connection and on their digital skills.

5.3 Item-writing skill acquisition as an individual cognitive process

5.3.1 The nature of item-writing skills

Findings from the present study, as discussed later in this chapter (Section 5.3.3), suggest that there exists no unitary item-writing skill to produce all types of item. Instead, different items involve partly different skills from item writers. Therefore, I use the term 'item-writing skills' in plural throughout this chapter. In psychology, skills are categorised into motor, basic, communication, social, and cognitive (American Psychological Association, n.d.). Cognitive skills/abilities[11] are "involved in performing the tasks associated with perception, learning, memory, understanding, awareness, reasoning, judgement, intuition, and language" (American Psychological Association, n.d.). It follows that the item-writing skill can be categorised as cognitive. I hypothesise that item-writing skills are acquired according to general principles of cognitive skill acquisition as described in ACT-R (Anderson, 1993). Namely, novice item writers first acquire *declarative* knowledge about item writing. This declarative knowledge is converted into *procedural* via item-writing practice. The process of declarative item-writing *knowledge compilation* (or *proceduralisation*) results in the formation of item-writing *production rules*, which can be defined as individual components of an item-writing skill. Newly formed production rules then go through the process of *tuning* which involves production rule *strengthening* whereby "better rules are strengthened and poorer are weakened" (Anderson, 1996, p.241) and experimentations with increasing production rule effectiveness. Finally, after a long period of professional practice, item writers might demonstrate commensurate expert performance (*Figure* 5-1). The process of item-writing skill acquisition is discussed in more detail in Section 5.3.2.

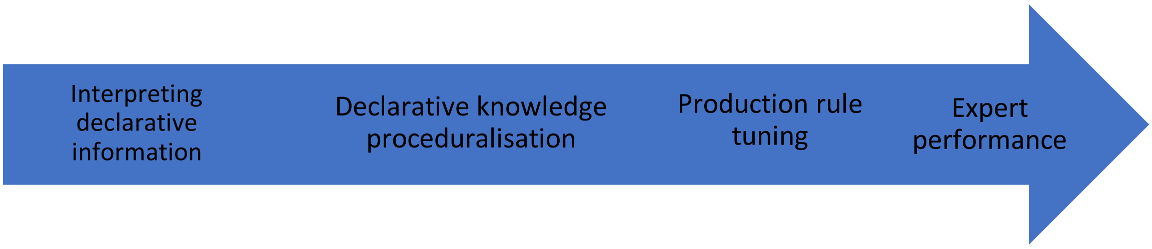

Interpreting declarative information
Declarative knowledge proceduralisation
Production rule tuning
Expert performance

Figure 5-1. The process of cognitive skill acquisition

It is widely acknowledged that item-writing skills are difficult to master (e.g., Buck, 2009; Shin, 2012). Partly, this might be because they consist of many components. The idea of a cognitive skill as a set of production rules is central to the ACT-R theory (Anderson, 1993) and received considerable attention in later psychology research; for example, Speelman and Kirsner's (2005) *Component Theory of Skill Acquisition* takes skill componentiality as its main premise. Findings from the present study support the idea that item writing is enabled through using a number of production rules. For instance, to write a listening task adhering to the specifications used in this study (*Appendix 4*), item writers should produce a text, a set of gap-fill items, task instructions, and a comprehensive answer key. To produce a text, for example,

[11] In psychology, the terms 'skill' and 'ability' are used interchangeably and are seen as different from 'capability' which is defined as "an ability, talent, or facility that a person can put to constructive use." (American Psychological Association, n.d.).

item writers need to be able to choose an appropriate topic and genre; decide on content that is inoffensive, culturally unbiased, and appropriate for use in a test; produce a text that sounds authentically spoken; consider what information to target in items; and skilfully incorporate the required number of distractors in the text. Item writers should also be able to keep the text within a certain word limit while meeting the vocabulary frequency and grammar level requirements. Item writers are supposed to be equally proficient at each of these to ensure that the listening task is acceptable for live testing.

This study's findings led to three observations with regard to the component nature of item-writing skills, as elaborated below: (1) item-writing production rules can be categorised into (more) objective and (more) subjective; (2) the formation of different item-writing production rules does not happen at the same pace; (3) during item writing, several production rules are executed (or *fired*, according to ACT-R terminology) simultaneously and the ability to produce high-quality items depends on the item-writer's ability to co-ordinate the execution of these rules, which might require a separate production rule.

5.3.1.1 Objective and subjective item-writing production rules

As explained in the Methodology Chapter (Section 3.5.1.4), for reasons of practicality, I myself scored the items produced by participants on all criteria that could be evaluated objectively, such as word count or vocabulary frequency. The independent reviewers evaluated the items on criteria requiring subjective judgement, such as the construct or strength/plausibility of distractors. For most objectively-scored criteria, large or even statistically significant improvement in item quality was found following the training. Improvement in quality on subjectively-scored criteria, however, was uneven among the trainee cohort; gain ratio statistics revealed a large post-course score gain for some trainees, smaller gains for others, and no gain or lower post-course scores than the corresponding pre-course ones for some.

The differing trends for objectively- and subjectively-scored criteria might be related to differences in the relevant production rules. I use the term '*objective item-writing production rules'* to refer to the rules which execution is directly measurable, such as producing a grammar MC item stem of up to 10 words. Compliance with the relevant specification requirements can be checked by item writers themselves using simple maths, existing tools (e.g. *Lextutor*, *Cohmetrix*, spell-checker) or documentation (e.g. the list of topics / functions in the *Core Inventory*). The execution of '*subjective item-writing production rules*', on the other hand, relies on creative ability, general writing ability, or knowledge of language assessment principles. Compliance with the relevant specification requirements cannot be measured directly but requires subjective judgement by both the item writer while producing an item and the item reviewer while deciding on item acceptability. For instance, to produce a writing prompt for the set of specifications used in this study (*Appendix 4*), item writers had to think creatively to come up with a plausible scenario that would be suitable for the test-taker population and elicit an original response; they had to write a prompt which would elicit certain writing (sub)skills from test-takers, thus requiring item-writer's knowledge of writing constructs; the prompt also had to be clear and well-written, something which calls for some general writing ability.

5.3.1.2 The pace of item-writing production rule formation

The predominantly fast and linear increase in scores on most objectively-scored criteria following the training might indicate that objective item-writing production rules can be formed and tuned relatively quickly with the help of an induction item-writing course that includes training in the use of item specifications, other related documentation, and item-writing tools (as was the case regarding the course in this study). Conversely, subjective item-writing production rules might require more time to form/tune; this study found no significant changes in the total post-training scores on subjectively-scored criteria, which was true for all item types. This suggests that the formation/tuning process of subjective production rules might need to continue beyond initial training.

This difference in pace is in line with the ACT-R theory which posits that some production rules can be formed after a single trial, while for others the process might be more gradual (Anderson, 1993c). In item writing, the reason for the slower subjective production rule formation/tuning might be that considerable cognitive effort is required from novice item writers to comply with the relevant specification requirements. This might be explained with less certainty in decision-making about what meets the requirements, thus the requirements are more demanding on cognitive resources. Moreover, complying with subjective requirements often involves considering a host of different factors, while complying with objective requirements is normally more straightforward. For example, indicating the MC grammar item key with an asterisk is straightforward, while making sure that the stem provides enough context to ensure that the intended construct is tested involves a host of considerations.

It also seems that a considerable number of objective item-writing production rules can be formed simultaneously, as item writers in this study were able to successfully meet most objective specification requirements at the post-course stage. However, the jagged post-training item score profiles found with regard to individual subjectively-scored criteria indicate that some subjective item-writing production rules might be formed faster while others might lag behind. One reason for this might be the importance a trainee attaches to a particular specification requirement. For instance, as revealed in the post-training interviews, many trainees in this study paid particular attention to grammar item distractors and listening text authenticity. As a result, the post-training scores on these criteria were statistically significantly higher compared to the pre-training ones. The way item-writing training is organised might influence what trainees perceive as more/less important and, consequently, pay more attention to while writing items. For example, this study's training in producing listening tasks included a lot of focus on input text authenticity. Not entirely unsurprisingly, trainees discussed text authenticity at length in post-training interviews, and – as hoped – post-training scores on the text authenticity criterion showed a statistically significant improvement.

The degree of subjectivity in subjectively-scored requirements could vary, which might be another contributing factor in the pace of the associated production rule formation. For example, the requirement for grammar item distractors to be grammatically correct as a stand-alone proved easier for trainees to master than the requirement for the distractors to be strong and plausible. The former requirement is less subjective because the compliance can be checked against existing grammar rules, while the latter is more subjective as it has no fixed reference point.

5.3.1.3 Co-ordinating item-writing production rules

This study found that item-writing production rules are not executed independently of each other; they are often fired simultaneously and have to be balanced against each other. For example, to produce a writing prompt for this study's assignment, participants had to balance the requirement to create an input message that includes sufficient information for eliciting a desired response, with the requirement to keep the input message at a certain length; or, grammar item distractors had to be grammatically correct as a stand-alone but incorrect within the stem. The relationship between different requirements can also form some tension, with the requirements seemingly being in competition with each other, for example, the need for an input listening text to sound authentically spoken but also to have in-text distractors which are not normally a feature of authentic discourse. This balancing act of conforming to all requirements simultaneously might call for a production rule different from the production rules responsible for conforming to individual specification requirements.

This study found that balancing competing requirements was difficult for participants even after the training, with post-training scores on some such criteria being lower than the corresponding pre-training ones. When one of the criteria was objectively-scored and the other subjectively-scored, an interesting tendency was found: pre-training, participants were better able to comply with the objectively-scored criterion requirement, while the items scored low on the related subjectively-scored one. After the training, however, the opposite was the case. This happened, for example, for the writing prompt input message word-count vs. the input message plausibility and clarity; and for the listening input text's word-count vs. the text's authenticity. It might be that, following the training which increased participants' awareness of the subjective requirements, the participants focussed on these requirements more in their post-training items. This took up their cognitive resources and so, because they were not yet able to balance both requirements, they became less attentive to the related objective requirement. However, it was also found that so-called 'high-achievers', who initially demonstrated better ability to produce items and whose item-writing skills developed further following the training, were often able to meet both competing requirements. This finding might indicate that there exists a production rule that allows an item writer to meet competing specification requirements by balancing the execution of the relevant production rules, and this production rule is formed later in the process of item-writing skill acquisition.

5.3.2 The process of item-writing skill acquisition

This study's findings revealed several participant profiles, the most prominent of which are 'outliers', 'high-achievers', and those whose items scored lower following the training. The trajectories of item-writing skill acquisition for the three profiles are discussed in the section that follows. There were also several participants whose items received similar scores on both occasions – they were not discussed separately in the Results Chapter and are not discussed in this chapter due to the necessity to limit the scope of this thesis. Overall, it seems that different participants walked the path of the item-writing skill acquisition in different ways, as discussed in Sections 5.3.2.1 -5.3.2.3. However, the path itself was the same, which is discussed in Section 5.3.2.4.

5.3.2.1 Profile A: The trajectory of item-writing skill acquisition for participants who produced ‘outlier’ items pre-training

There was a small number of items for each item type that scored much lower than the rest before the training. In the Results Chapter, these items were characterised as ‘outliers’. All profile A participants wrote higher-quality items following the training.

The pre-training item-writing approach of these participants, as revealed in their interviews, displayed some common characteristics. Firstly, they paid secondary or no attention to the specifications; those who attempted to attend to the specifications found them difficult to understand. At the same time, they treated examples provided with the specifications as models to copy indiscriminately in their items. This is in line with Kim et al. (2010) who found that inexperienced item writers in their study did not like reading the specifications and did not rely on them during the item-writing process, while they perceived example items as more useful. Unfortunately, the inferences about item requirements that profile A participants made from studying the example items were sometimes wrong, which is in line with the ACT-R theory positing that *analogy compilation* (i.e. learning by examples) might sometimes lead to misinterpretations and “mistaken inferences” (Anderson, 1993b, p.88).

Secondly, the specification requirements that profile A participants focussed on were mostly objective, arguably because those are more concrete and, therefore, easier to understand. They overlooked or misunderstood many subjective specification requirements, which was probably the result of not attending to/not understanding the specifications. At least some of these participants also misunderstood some of the objective requirements or were unable to use the relevant item-writing tools and documentation.

The training in using the specifications and the item-writing tools was provided during the course; it was effective in developing these trainees’ ability to interpret the specifications and to comply with objective requirements. It also helped to adjust the approach of those participants who had their own ideas, probably based on their previous experience, of what was required for a particular item. For example, pre-training, Ted thought that a writing prompt had to be entertaining while post-training he aimed to elicit a required response. At the same time, most items produced by these participants after the training still required further revision, which is especially true for the listening tasks and for subjective specification requirements of all tasks.

Using ACT-R terminology, these trainees encountered difficulties with interpreting the declarative item-writing information (that is specifications, example items, and instructions on using item-writing tools) that they received for the pre-training assignment. Unlike participants in profile B and C groups, they were unable to independently interpret this information and required explicit instruction. Because these participants took longer to acquire the declarative knowledge, the process of *proceduralisation* might have started later for them, and they still largely relied on declarative knowledge while producing items post-training. This is particularly true for subjective specification requirements because, as discussed in Section 5.3.1.1, subjective production rules might take longer to form. At the same time, because objective production rules might be faster to form, these participants performed markedly better in complying with objective requirements after the training. Profile A participants might need more item-writing practice for subjective production rules to form, as well as for production rule tuning to happen (*Figure 5-2*).

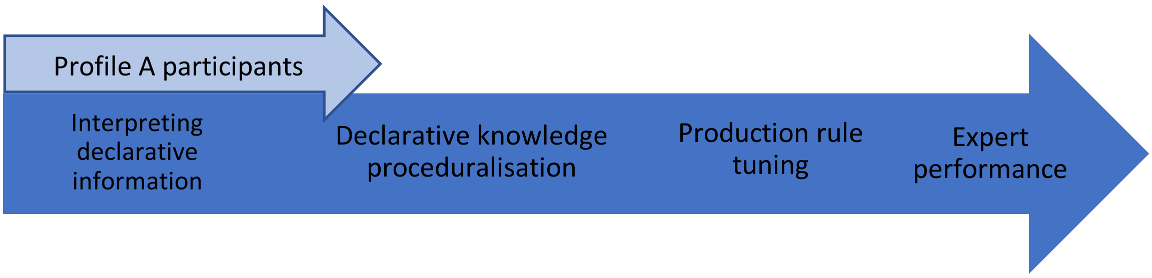

Profile A participants
Interpreting declarative information
Declarative knowledge proceduralisation
Production rule tuning
Expert performance

Figure 5-2. The trajectory of item-writing skill acquisition for profile A participants

5.3.2.2 Profile B: The trajectory of item-writing skill acquisition for 'high-achievers'

A considerable number of participants produced relatively good-quality items before the training and delivered even better-quality items post-training, scoring the maximum or near-maximum. There were also several participants whose item scored the maximum on both occasions. Such participants were characterised in the Results Chapter as 'high-achievers'.

Pre-training items produced by profile B participants scored substantially higher than those written by profile A participants, with no band '0' and a number of band '2' scores awarded for individual criteria. High-achievers' pre-training interviews revealed that they were generally able to understand the specifications and use the item-writing tools, they also recognised the importance of following the specifications and did not overestimate the role of example items. However, because these participants relied on declarative knowledge in producing the items, and declarative knowledge consumes a lot of memory capacity (Anderson, 1996), they might have been unable to keep the whole set of specifications in their working memory. This might explain why the participants sometimes had to 'retro-fit', that is amend already-written items, having noticed that a requirement had been overlooked. 'High-achievers' were also generally able to meet the objective requirements, and they did not focus on those excessively in interviews; instead, they discussed the subjective requirements and demonstrated some awareness of test constructs and related issues.

After the training, these participants demonstrated more efficient use of the specifications. For example, they took time to study the specifications before the item-writing event, they referred to the specifications repeatedly while writing items and, having produced an item, they checked it with the specifications again (see Section 4.3.1.3). Notably, they did not normally have to 'retro-fit' post-training. This might point to better familiarity with the specifications, but also to having more cognitive resources available because at least part of the item-writing knowledge had become proceduralised, thus freeing up working memory. Importantly, the constant attention to the specifications is in contrast with expert item writers who, as reported by Salisbury (2005) and Green and Hawkey (2011), relatively rarely consulted the document.

Post-training, these participants also reported more effective ways of producing items, compared to how they produced the items pre-training. One example was introducing the pre-writing stage to the item-writing process, which helped eliminate false starts and ensured that fewer attempts at the item were needed, thus saving time. For example, Logan used vocabulary frequency lists to select the writing prompt scenario, while Nathan brainstormed ideas for the listening task – something that resonates with Salisbury (2005) whose expert writers of listening tests spent "a great deal of effort identifying an effective context, in the knowledge that this will allow rapid subsequent test instantiation" (p.289). The pre-writing stage is something that was also described for experienced item writers in Fulkerson and Nichols (2010) and Johnson et al. (2017). 'High-achievers' also found more effective approaches to producing particular types of item. For example, they reported writing the text and items for the listening task simultaneously and iteratively, and many of them also produced the listening input texts through semi-scripting/vocalization.

Post-training, 'high-achievers' seemed to require no effort in conforming to objective requirements (see Sections 4.3.2.1 – 4.3.2.3). As a result, their working memory became freed-up in order to attend to subjective requirements, resulting in a deeper and more nuanced consideration of such issues as construct, authenticity, and distractor strength and plausibility. For example, Mason's discussion of the writing prompt construct was much clearer and more coherent in the post-course interview. This deeper understanding might also be linked to the development of language assessment literacy which received a lot of attention during the training (see Section 3.2). Finally, ‘high-achievers’ were more successful at balancing competing specification requirements post-training because in many instances their items scored band ‘2’ on the relevant criteria.

Although high-achievers’ post-training items were of generally good quality, many of the items still required minor revisions to be accepted for live testing. The participants themselves felt they had not yet fully mastered the skill of item-writing, as is clear from their post-training interviews. Importantly, those who were discussing the need for further skill development, did not request more training but said that they needed more item-writing practice and feedback.

Using ACT-R terminology (*Figure 5-3*), these trainees were successful at independently interpreting declarative item-writing information at the pre-training stage. One explanation might be that they were more invested in testing when doing classroom assessment and/or acting as examiners, so they paid more attention to assessment-related information when they encountered it in their professional practice, which means that they had more background knowledge to rely on when interpreting the declarative item-writing information. Their digital literacy might have played some role too – all item-writing materials and tools were located online. However, these are only speculations, as explaining why participants with similar backgrounds differed in acquiring item-writing knowledge goes beyond this study’s aims and requires research into individual participant characteristics.

Faster declarative knowledge acquisition allowed for an early start in production rule formation. For ‘high-achievers’ some production rules – especially objective ones – might have formed on one trial when producing pre-training items. Because of this, high-achievers’ working-memory capacity was freed-up to attend to subjective specification requirements during the training, and many subjective production rules were formed during or following the course. Moreover, post-training, these participants reported more effective ways of producing items, which points to production rule tuning. These participants were also largely successful at meeting competing requirements, which suggests that the production rules responsible for balancing such requirements had also been formed. Due to these, ‘high-achievers’ might have advanced further on the path of item-writing skill acquisition compared to profile A participants. However, mastery had not yet been achieved, which is in line with skill acquisition research suggesting that typically years of deliberate practice are needed to reach expert status (Proctor & Dutta, 1995).

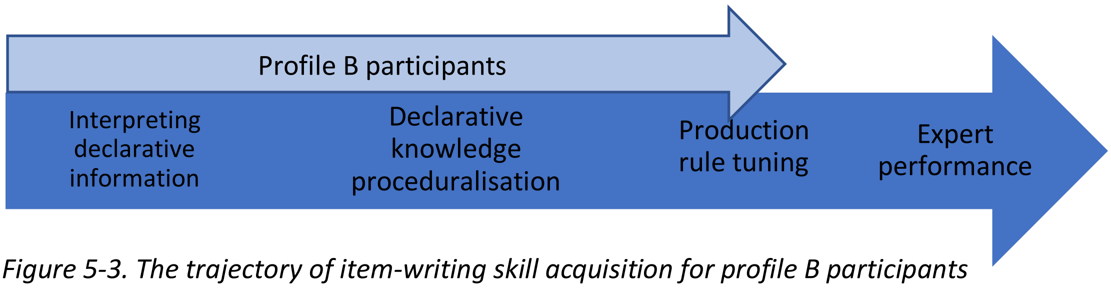

Profile B participants
Interpreting declarative information
Declarative knowledge proceduralisation
Production rule tuning
Expert performance

Figure 5-3. The trajectory of item-writing skill acquisition for profile B participants

5.3.2.3 Profile C: The trajectory of item-writing skill acquisition for those participants whose items received lower scores post-training

A number of participants produced items for the post-training assignment that scored lower than the corresponding pre-training ones. These participants shared two common characteristics: (1) their pre-training items were of generally good quality; (2) the decrease in scores post-training was by one or few points only. These participants' pre-training scores indicated that, similar to profile B trainees, they were successful at interpreting declarative item-writing information before the training; their post-training interviews also contain many features in common with the interviews of profile B participants. For instance, they demonstrated deeper understanding of the specifications and testing constructs, they largely focussed on subjective requirements, they also reported some "trial-and-error exploration" (Anderson, 1996, p.241) characteristic of production rule tuning.

Analysis of these participants' post-training interviews revealed three potential reasons for the decrements in the post-training performance, which is in line with Strauss and Stavy (1982) who suggested multiple reasons for U-shaped learning curves.

1) Excessive focus on a particular aspect of item-writing

The most common reason for U-shaped learning was the participants' preoccupation with a particular aspect of item-writing after the training, while some other aspect(s) might have got overlooked. The specification requirements that these participants paid heightened attention to were always subjective ones, while the requirements they overlooked were either subjective or, less frequently, objective (when it was a part of a subjective-objective competing pair). For instance, listening text authenticity occupied much of the trainees' attention post-training, while the text word limit became overlooked to the extent that the scores on this criterion were statistically significantly lower compared to pre-training. There exists an empirically proven (Carter & McCarthy, 1997; Gilmore, 2004) relationship between these two requirements: it was found that genuine spoken texts are normally longer than scripted ones because of spoken language features such as filled pauses, hesitations, reformulations, false starts, and so on (see a discussion in Rossi & Brunfaut, forthcoming). It appears that, prior to the training, these participants were able to keep to the word-limit but unable to make the text sound authentic. After the training, the opposite was true. It would be wrong to assume that, having developed the ability to produce authentic-sounding texts, the participants simultaneously lost the ability to produce texts of a certain length. Rather, they had not yet acquired the ability to co-ordinate the production rules responsible for meeting the two requirements. This explanation finds support in Strauss and Stavy (1982) who wrote that one reason for a U-turn might be acquiring all components of a complex skill but not being able to co-ordinate those components.

The question arises why these participants focused on one item aspect and not some other. The focus might have been provoked by something that grabbed the participants' attention during the training. A case in point is the listening text authenticity requirement. Pre-training, the participants were generally unable to produce authentic-sounding texts. Having learnt the features of authentic spoken language during the course (which was new to many of them), the participants felt enthusiastic in producing authentic-sounding texts post-training, as their interviews revealed. Another trigger could have been the feedback the participants received on their pre-training items: some participants mentioned in the interviews that they revisited the feedback before writing their post-training items. Because the feedback focused on the areas of weakness, it could not serve to reinforce the strong points of the items. Subsequently, the requirements that tended to be overlooked by these participants in the post-training items were the ones they received high scores for before the training. This finding suggests that item-writing trainees should receive feedback on their items' areas of strength and not only of weakness: negative feedback aims to eliminate unhelpful item-writing habits, while positive feedback helps to reinforce helpful ones.

2) Experimentation with item-writing approaches

Post-training interviews revealed that some of the profile C participants used an approach that did not prove effective, while they used a more effective approach pre-training. Lucas' experimentation with producing listening tasks can serve as an example. Before the training, he first wrote items and then the text. After the training, he reversed the order, which resulted in lower scores on four item-related criteria. This is because, as found in this study and also by Salisbury (2005), producing a listening text without giving consideration to the items might result in subsequent difficulties with producing items – the items might be unclear and/or might not target the intended construct; moreover, the text will most probably lack the required distractors, while retro-fitting the distractors into the text will make them too conspicuous and, therefore, weak.

Experimentations with item production, as was the case for Lucas and several other profile C participants, is a sign of production rule tuning (Anderson, 1996). It seems that these participants were looking to increase production rule effectiveness but the modifications they introduced failed to work. Similar to what was discussed above, the experimentation might have been triggered through training input. The reason for a U-turn in their case is similar to what was described by Strauss and Stavy (1982): the learner has two production rules – a familiar but inadequate one and a new but 'untrusted' one. This failure in production rule tuning might trigger the second step in the learning cycle as described by Hayes-Roth et al. (1981) whereby learners "diagnose the problems in behaviour and refine the knowledge that underlies them" (p.233). However, more longitudinal research would be required to check this hypothesis.

3) Exploring the boundaries of item specifications

One more reason for the U-turn might have been some participants' desire, after the training, to explore the boundaries of the specifications by writing something substantially different from the example item. The desire might have stemmed from the participants' increased confidence in writing items, improved awareness of the role of the specifications, and the realisation that the example item was not the only way to operationalise the specifications. The writing prompts produced by Adam, Arthur and Joe might serve as examples. The example prompt in the specifications had an email as the input message, and the three participants successfully produced email input messages before the training. After the training, however, they wanted to do something different from what they did before, so they decided to produce a public notice. However, it seems that the participants, despite being native/proficient users of the language, were not aware of the genre conventions of public notices, which resulted in lower scores on several input-message criteria. As noted by Gilmore (2015), "native speaker intuitions about language and speech behaviour are notoriously unreliable" (p.515). This might point to the need to train item writers in producing texts in each genre included in the specifications. In operational settings, item writers might also benefit from being provided with a range of example items that reflect the breadth of the specification requirements.

Importantly, the U-shaped learning curve was not observed for profile A and B participants. This is in line with research into the development of other cognitive skills, for example child acquisition of maths (Strauss & Stavy, 1982), where the U-turn was also not found to be universal. McLaughlin (1990) wrote that "practice can have two very different effects. It can lead to improvement in performance as sub-skills become automated, but it is also possible for increased practice to lead to restructuring [i.e. a U-turn] and attendant decrements in performance as learners reorganise their internal representational framework" (p.125). It seems the effect of practice on profile C participants was such that it provoked a U-turn, while practice led to observable linear improvement in the performance of profile A and B participants.

Neither Strauss and Stavy (1982) nor McLaughlin (1990) provided an explanation of why practice might have different effects on learners. With regard to item-writing skill acquisition, I hypothesize that at least one of the contributing factors might be the trainees' working memory capacity. Item writers with larger working memory capacity might be better able to attend to all requirements simultaneously, which helps mitigate the effect of paying excessive attention to a particular requirement the participants want to experiment with. This suggestion finds support in Salisbury (2005) who wrote that "efficient aural memory – both working and long-term" (p.293) is a pre-existing ability that benefits writers of listening tasks.

To sum up the profile C participants' trajectory of item-writing skill acquisition using ACT-R terminology (*Figure 5-4),* these participants were successful at interpreting declarative item-writing information at the pre-training stage. This allowed for (some) production rules to be formed, which made their initial skill acquisition fast. However, production rule tuning which was happening after the training resulted in some poor choices, leading to a U-turn in performance. Moreover, these participants were also less able to co-ordinate competing requirements, pointing to the relevant production rules not having been formed.

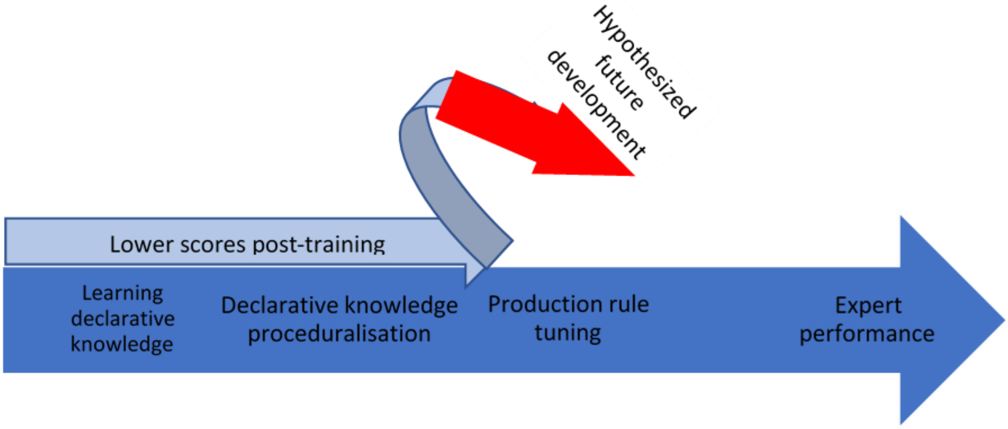

Hypothesized future development
Lower scores post-training
Learning declarative knowledge
Declarative knowledge proceduralisation
Production rule tuning
Expert performance

Figure 5-4. The trajectory of item-writing skill acquisition for profile C participants

5.3.2.4 The process of item-writing skill acquisition

The three trajectories of item-writing skill acquisition discussed above support the suggestion that, although individual item writers might walk the path of item-writing skill acquisition in somewhat different ways and at a different pace, the path itself is actually the same.

This study's findings demonstrated that item-writing skill acquisition starts with interpreting declarative item-writing information. The information can come in different forms: in this study, participants were provided with item specifications, example items, and additional item-writing documentation and tools (the *Core Inventory*, *Lextutor*) for the pre-training assignment. The various types of training input that participants received during the course (e.g., language assessment principles, tips on producing items of different type) can also be considered declarative information. Although the information was the same, participants seemed to differ in their ability to interpret it: participants in B and C profile groups were able to independently interpret the instructions they received for the pre-training assignment, which led to higher evaluations of their pre-training items. Profile A participants, on the other hand, experienced difficulties with interpreting the information, consequently, their pre-training items were of lower quality. Initially, all participants were guided by example items, which is in line with ACT-R which posits that analogy compilation is the most common route of skill acquisition (Anderson et al., 1997). However, it seems that more successful participants, although they used example items to guide them, paid equal attention to the specifications, as opposed to profile A participants, who were largely unable to interpret the specifications.

Declarative information interpretation is followed with the formation of production rules. This study's findings suggest that item-writing production rules were formed at a different pace by different participants. For those in the B and C profile groups, at least some of the production rules seemed to have been formed on their first item-writing attempt, which is in line with ACT-R which posits that "a production rule can be created after a single example" (Anderson, 1993, p.87). This is particularly true for objective production rules which might require less practice to form. For profile A participants, however, the formation of production rules was delayed due to the failure of declarative information interpretation. For these participants some production rules might have been formed during the training, after the declarative information had been clarified, while many production rules were still unformed, or in the process of formation, following the training.

Having been formed, item-writing production rules need tuning, whereby effective rules are strengthened, ineffective rules are discarded, and modifications are made to existing production rules to make them more effective. Production rule tuning is characterised with less attention to example items and more focus on specifications, which is in line with ACT-R's positing that example-based processing is gradually replaced with rule-based processing. "Trial-and-error exploration" (Anderson, 1996, p.241) is the main feature of tuning, and exploration of specifications can facilitate this process because multiple item variations can result from the same set of specifications.

It seems that profile B and C participants reached the production tuning stage, which happened later in the training or when the participants were producing items for the post-training assignment. However, the process of tuning differed for these two participant groups. The tuning was more successful for profile B participants whose explorations led to more effective item-writing approaches and higher item evaluations. For profile C participants, the explorations often resulted in failures and, consequently, lower scores on their post-training items. One reason for such a difference, as hypothesised in Section 5.3.2.3, might be working memory capacity: profile B participants' working memory capacity might be larger, therefore production rule tuning, which draws on item-writer's cognitive resources, did not result in overlooking other aspects of the item, something that happened for profile C participants. Moreover, there was another difference between profile B and C participants following the training – profile B participants had a better ability to balance competing specification requirements.

Finally, this study demonstrated that none of the participants, even the most successful ones, walked the path of item-writing skill acquisition to the end following the training, that is achieved full item-writing mastery. This is in line with the current research into skill acquisition – it was found that many years of consistent practice are needed to reach the expert status in a particular domain (Proctor & Dutta, 1995). Based on the ACT-R theory of skill acquisition, I hypothesise that expert item-writing performance would have the following characteristics: (1) full, accurate internalisation of all item-writing requirements for a particular item type, as well as all other relevant item-writing information, including language assessment principles underlying the production of a particular item type; (2) strong, fully-formed production rules that allow for confident item production resulting in high-quality items; (3) mastery in complying with competing specification requirements; (4) the production rules are well-tuned, which means that the item writer uses most effective approaches resulting in relatively fast performance. However, because of the complexity of the process and the large amount of production rules involved, item writing might be a slow and labour-intensive activity even for most expert item writers.

To sum up, the process of item-writing skill acquisition seems to happen as follows: first, new item writers are presented with item-writing declarative information which they need to interpret. Following the interpretation, which might or might not require explicit instruction, item-writing production rules start to form. Some of them, in particular the production rules responsible for complying with objective specification requirements, are formed fast, often on the first item-writing attempt; other rules, in particular the ones responsible for complying with subjective specification requirements, might take more item-writing practice to form. The production rules responsible for balancing competing specification requirements are formed after the rules responsible for complying with each individual competing requirement. Having been formed, production rules are tuned. For item-writing, this means discovering the production rules that work best and discarding others, as well as introducing modifications to existing production rules to make them more effective. Tuning failures slow down the process of item-writing skill acquisition: they require problem diagnosis and, possibly, further clarification of declarative information. The diagnosis and clarification might be provided during the induction training but, because tuning often happens following the training, the clarification might have to come in the form of reviewer feedback, or item writers might receive further training while in employment.

Importantly, item-writing production rules are not formed and tuned simultaneously, which might mean that novice item writers use a mix of declarative and procedural knowledge while performing an item-writing task, which is in line with the ACT-R theory (Anderson, 1996). The objective production rules are often formed first, subjective production rules might take longer to form, while the production rules responsible for balancing competing specification requirements might take the longest. This observation means that it is impossible to tell with confidence which stage of item-writing skill acquisition a particular item writer is at – it depends on the item writer, the item type, as well as on individual production rules for each item type. It is also difficult to predict how long item-writing skills might take to develop: the initial declarative information interpretation might happen fast or might take longer, but even if an item writer moves fast initially, s/he might experience a U-turn later when production rules are being tuned. This is in line with what was found by Anderson et al. (1993a) when researching skill acquisition for solving geometry problems: "students differed not only in their initial ability but also in their learning rate" (p.179).

5.3.3 Item-writing skill acquisition for items of different type and proficiency level

The present study empirically confirmed Wesman's (1971) observation that item writing is not a generic skill but is at least partly item-type specific. All participants displayed jagged profiles with, for example, one and the same trainee being an 'outlier' for the grammar items, a 'high-achiever' for the writing prompt and having a decrement in performance for the listening task. Moreover, even for the same item type – MC grammar items – some participants demonstrated a different trajectory of skill development for A2 and C1 items. I therefore hypothesise that, rather than acquiring a 'universal' item-writing skill, item writers acquire the skills for producing a particular item type, at a particular proficiency level. This might explain anecdotal evidence of item-writer specialisation, with testing organisations having preferences regarding whom to allocate item-writing commissions to.

Findings from this study showed that writing MC grammar items might be susceptible to faster skill development compared to the other two item types, with participants also saying that it was much easier for them to write grammar items following the training. However, there were also two areas of difficulty – targeting the intended construct and creating strong and plausible distracting options (see Section 4.3.2.1). The formation of relevant production rules might have something to do with the item-writer grammar knowledge and might be item proficiency-level specific. The majority of participants in this study were native speakers of

English who, in interviews, professed a lack of explicit grammar knowledge despite being qualified and experienced teachers of English. The mean score on the C1 item construct (*wish/if only* to express present and past regrets) was low pre-training but increased considerably following the training. This might be because many participants held the belief that higher-level grammar items were harder to write and, being aware of their insufficient grammar knowledge, put more effort into learning to target the C1 construct during the training. On the other hand, the mean score on the A2 item construct (*wh*-questions in the past) did not increase from before to after the training. Lower-proficiency items were (wrongly) perceived as 'easy' to write by participants, so they might not have given sufficient care in learning to target the A2 construct in items.

Producing strong and plausible distractors saw the opposite trend: for C1 items, the pre- and post-training mean scores were similar, while for A2 items the post-training mean score was statistically significantly higher than the pre-training one, indicating that the participants' ability to produce strong and plausible distractors for A2 grammar items improved, while for C1 items it did not. During the course, participants were introduced to some general principles of creating good distractor options. Because A2 distractors were generally shorter and less complex, it might have been easier for the participants to successfully apply the principles in A2 items. On the other hand, for C1 distractors to be strong, they must reflect the complexity of the C1 construct – something that many novice item writers might have failed to consider. Notably, awareness of the necessity to pitch distractors at the right level of proficiency was something that only 'high-achievers' demonstrated in interviews (see Section 4.3.2.1).

Findings from the writing prompts' evaluations suggest that participants' initial ability to produce writing prompts was higher than their initial ability to produce listening tasks. This might be related to participants' previous experience seeing writing prompts as examiners and/or producing writing tasks for classroom assessment, while they might have had much less experience producing listening tasks. This is because, in my experience, teachers often create their own writing/speaking prompts while they prefer to use ready-made listening and reading tasks, probably because of the difficulty and time required for the latter. Moreover, the writing prompt specifications were considerably less complex (16 evaluation criteria compared to 26 for the listening task). However, the skill of producing writing prompts also showed the least development, judging from the post-course writing prompt evaluations. This might be, in part, related to participants' perceptions. Most participants said both pre- and post-training that writing prompts were the easiest to produce. This perception might have led to less effort in learning to produce writing prompts, with more nuanced requirements, such as prompt plausibility and clarity, not having been given sufficient consideration, which resulted in no score increase on the relevant criteria following the training. Further training and more item-writing practice might be needed for novice item writers to master more nuanced aspects of writing prompt production.

However, little improvement in the post-course writing prompts' evaluation scores might also have to do with factors unrelated to the item-writing skill acquisition. Firstly, as discussed in the previous paragraph, pre-course writing prompts were already of reasonable quality, which made the post-course increase in scores less perceptible. Secondly, there were much fewer objectively-scored criteria for writing prompts compared to grammar items and listening tasks (five, ten and twelve, respectively). Wilcoxon signed-rank tests demonstrated that it was the increase in scores on objectively-scored criteria that was largely responsible for the significant increase in the total item scores with respect to grammar and listening items. Because writing prompts were evaluated on much fewer objectively-scored criteria, the improved scores on these criteria did not make a large enough contribution to the total score for the change to be statistically significant.

The listening tasks were the most challenging for participants to produce initially and, although there was a statistically significant increase in the post-training total scores, the listening tasks were still awarded the lowest scores among the three item types following the training. This is unsurprising, given that many more issues and features need to be considered when producing listening tasks, compared to MC grammar items or writing prompts. Post-training, participants said that the listening task was the most difficult for them to write because there was "*an awful lot to* ***keep in your head***" (Ted), indicating that many production rules were not yet formed and the item-writing knowledge was still used in its declarative form. Because of the large amount of declarative information involved, the difficulty of paying attention to everything at once could have been overwhelming for many participants. I hypothesize that item writers might continue producing relatively low-quality listening tasks until some of the component production rules are formed. Following that, skill development might happen faster because working memory will be freed up to concentrate on the most challenging item aspects.

This study's findings suggest that different item-writing production rules have different rates of formation. Objective production rules were the fastest to form, while subjective production rules might take longer. Moreover, the rates of formation for the latter were not homogeneous. For example, it seems that the rule for producing authentic-like input is susceptible to faster formation, while the rule of producing strong distractors takes longer to form. This might be related to the nature of the underlying knowledge and the amount of learning involved. Although the concept of authenticity was new to participants, they could relate it to their everyday language experience; therefore, they only required the features of spoken language to be highlighted to them to start noticing such features and producing authentic-sounding listening texts (Rossi & Brunfaut, in press). On the other hand, participants could not rely on their previous language experience with regard to distractors. This might explain the slower rate of development, although distractors were afforded a similar – or even larger – amount of attention during the course. Even after the training, some participants were not fully clear about creating distractors, as discussed in interviews; in future more item-writing practice might be needed to internalise the concept and to form a production rule.

Notably, the requirements for items to be suitable for testing, culturally unbiased, not sensitive – all fairness-related requirements - were successfully met by many participants before the training, while the training must have led to the formation of suitable production rules because the relevant post-training mean scores were near-maximum or maximum, an observation which applies to all three item types. According to the ACT-R theory, this is because the production rules enabling fairness-related requirements were the same for the three item types, which allowed for positive transfer (Anderson & Singley, 1993). At the same time, participants had considerable difficulties in producing distractors for listening tasks, having previously practised producing distractors for MC grammar items. This might be because the production rules for producing MC grammar item distractors and listening gap-fill task distractors were different, which resulted in zero transfer.

Using Speelman and Kirsner's (2005) terminology, fairness-related production rules can be categorised as general, while distractor-related production rules are specific. Overall, it seems that most subjective production rules, with the exception of fairness-related ones, might be item-specific. On the other hand, objective production rules might be general, which is one of the reasons why they were formed faster than the subjective ones. For example, once an item writer acquires the habit of proofreading his/her items prior to submission, the habit applies to any type of item. The same might be true for the word-limit, topic, function, and vocabulary frequency requirements. It should be noted, though, that if an objective production rule is responsible for one requirement in a pair of competing requirements (e.g. word limit vs text authenticity for listening tasks), the transfer might not be possible because of interference from the subjective production rule which is item-specific.

5.3.4 The role of induction training in item-writing skill acquisition

5.3.4.1 The effect of induction training on participants in different profile groups

The findings demonstrated that the induction item-writing training in this study had different effects on individual participants, which might depend on the participants' initial item-writing ability and their individual characteristics, such as working memory capacity. There might also be other contributing factors that influenced learning, such as receptiveness and motivation.

It seems that the largest benefit from the induction training for profile A participants was learning to use item specifications, item-writing documentation and item-writing tools (see Section 4.3.2). They experienced initial difficulties with declarative item-writing information interpretation, and the course assisted them in clarifying specification requirements and using item-writing tools. The training also clarified some aspects of item formats they were unfamiliar with. As a result, the induction training provided these participants with an equal opportunity to start their item-writing career alongside participants in the profile B and C groups who demonstrated better initial item-writing ability.

Participants in profile B and C groups, on the other hand, did not seem to require training in using item-writing tools and documentation – they were able to understand how to use them by using them. This finding helps to explain how some of those item writers who were never formally trained still managed to acquire item-writing skills. However, some aspiring item writers, as the experience of profile A participants in this study demonstrated, might be unable to teach themselves item-writing and will require formal induction training for their item-writing skills to start developing. The lack of formal training when it is needed (Alderson, 2010) might explain the high drop-out rates among novice item writers, known anecdotally, and might be one reason why professional item writers are in short supply (Buck, 2009).

Although participants in profile B and C groups did not require training to interpret declarative item-writing information, there were other ways the training proved helpful for them. As discussed in Section 5.3.2.2, these participants formed some production rules on the first attempt before training, which allowed for the rules to start tuning during the training. The tuning process might have been facilitated by the input on language assessment principles underlying the whole item-writing practice. For example, the participants deepened their understanding of item constructs and made it more nuanced, which helped tune the production rules responsible for targeting the intended construct, resulting in higher post-training item evaluations on construct-related criteria. Moreover, the training provided these participants with some suggestions on how to increase production rule effectiveness, for example by using semi-scripting to produce listening input texts (Buck, 2001). This is in line with the ACT-R recommendation that training should introduce learners to "more powerful ways to solve problems" (Anderson & Corbett, 1993, p.238).

Judging solely from the post-training item evaluation scores, one might assume that the training had no beneficial effect on profile C participants. This impression, however, might be somewhat simplistic. These participants' post-training interviews demonstrated that their declarative item-writing knowledge had developed compared to pre-training, while their experimentations with item-writing approaches might be evidence of production rule tuning. The difference between these participants and the ones in the profile B group is that they were less successful in applying the training input to item-writing practice. It might be that, in line with Hayes-Roth et al.'s (1981) hypothesis, profile C participants required a further cycle of tuning. The cycle would have to happen outside of the induction training, however, for example through item reviewer feedback or in-service item-writing training.

The observation that the training resulted in different learning for participants in different profile groups suggests that induction item-writing courses should be designed to suit alternative trajectories of item-writing skill acquisition by providing a range of input, such as instructions in using item-writing tools/documentation, input on the principles of language assessment underpinning the item-writing practice, as well as suggestions on improving item-writing production rule effectiveness. Unfortunately, the latter might prove problematic. As stated in ACT-R, any skill training should be based on a thorough understanding of production rules that represent this skill. For item-writing skills this is, unfortunately, impossible due to the lack of research into item-writing, including item-writing processes and features of expert item-writing performance. Therefore, although the training in its current form proved effective for many participants, the training effectiveness could have been higher if more was known about item-writing production rules, how they are formed, and how to tune them.

5.3.4.2 A discussion of training features: Implications for item-writing training

Participants' feedback revealed that information about the principles of language assessment improved their understanding of testing constructs and provided a rationale for the inclusion of specific requirements in item specifications. Participants also said that the CEFR-related input helped them better target their items at particular proficiency levels. Notably, although the theory was generally perceived as useful and necessary, academic-style input (e.g. articles and book chapters) were perceived as less engaging. This might point to the need of providing theory in more accessible form for this audience, for example as a brief presentation rather than as an academic text. However, several participants enjoyed the academic readings and asked for a list of additional literature. This is another indication that the training should aim to serve diverse types of trainees, and one way of achieving this is by including optional readings and tasks.

One of this course's successes was in training participants to produce authentic-sounding listening texts. This was achieved through training in spoken language features and text production techniques as well as through practice in producing authentic-sounding texts. To the best of my knowledge, such training is not often a feature of item-writing courses but, judging from this training's results, it can be recommended.

This study revealed that, prior to the training, not all trainees had the ability to produce texts of the required genres, and not all of them had the knowledge of grammar which might be expected of them. The item-writing training in this study, as well as other item-writing training I know of (see, e.g., de Jong, 2008; Ingham, 2008) did not aim to develop the above knowledge and abilities, most probably for practicality reasons. The solution might be in either making the recruitment more stringent by, for example, assessing applicants' ability to produce well-written texts in required genres, or by introducing item-writer specialisation, as is already the case for some organisations whereby those participants who demonstrate better ability to produce texts in a particular genre receive further training and are then prioritised for relevant item commissions.

Participants' feedback indicated that the practical nature of the training, whereby participants regularly produced items and received feedback on these, was seen as the best feature of the course. This finding is in line with Salisbury (2005) who found that "the training item writers receive through feedback and discussion is often highly developmental" (p.75). Feedback on performance is also an important feature of skill acquisition in ACT-R (Anderson & Corbett, 1993). In the present study, individual feedback by the tutor was regarded as most beneficial, but feedback summaries oriented to the group as a whole were also perceived as useful. Because the summaries discussed typical item flaws and highlighted salient features of high-quality items, they might serve as an addition to item-writing guidelines; this was recognised by some participants. However, tutors might need to emphasize the feedback summaries' usefulness to ensure all trainees pay sufficient attention to them.

The course structure received extremely high evaluations from participants. Therefore, the following features, which characterised the course in this study, can be recommended for adoption in item-writing training: following the logical progression from theory to practice; offering a balanced combination of input, group discussions, and item-writing practice with feedback; having a similar structure to each training module; sequencing the input of declarative item-writing information in a way that allows information chunks to build on each other and to be re-used later in the training. The latter suggestion finds support in Speelman and Kirsner (2005) who argued for prioritising the development of general production rules that can be applied to a greater variety of individual tasks. The fact that this training included a variety of item types helped the production rule transfer, which might suggest that it is not recommendable for item-writing training to limit itself to one item type only as item comparison might help with "abstraction of features that are common to many items" (Speelman & Kirsner, 2005, p.74).

In their feedback responses, participants made suggestions for what might make item-writing training even more beneficial. For example, a number of participants asked for explicit training on item-writing processes for different item types. This could be done through presentations where the tutor "*talk[s] through their mental process of creating an item*" (anonymous feedback) and through item-writing guidelines that provide details of the item-writing process. Kim et al. (2010) also highlighted the need for training item writers in "organic [item-writing] principles that reflect their trial-and-error process" (p.165) and not only in Dos and Don'ts for creating particular item types. This seems to support Anderson and Corbett's (1993) suggestion that learners should be taught "more powerful ways to solve problems" (p.238). For example, findings from the present study suggest that producing/planning items before producing the text results in higher-quality listening tasks. The study by Salisbury (2005) also found that the item-first approach was more beneficial. Salisbury suggested that, once the item-first approach is adopted, "a whole sequence of alternative performance processes are possible [sic], leading to an items-first approach" (p.293).

It follows from these findings that item-writing training might accelerate the process of item-writing skill acquisition by facilitating production rule tuning, something that seems to take longer if novice item writers are left to their own devices to discover most effective item-writing approaches. For example, trainees can be taught to work with specifications using two screens (or split screens) or write out individual specification requirements as bullet points – something that only 'high-achievers' in this study did. In terms of individual item types, it seems beneficial to use common student mistakes to produce MC grammar item distractors, or to check writing prompt ideas for vocabulary frequency *before* producing the prompt. To help trainees with production rule tuning, the most effective item-writing approaches should be known to the trainers, which suggests that the item-writing process of expert item writers should be carefully researched.

It might also be beneficial to provide trainees with detailed positive, as well as negative, feedback on items because positive feedback might help in strengthening useful production rules. Repeated cycles of feedback were also requested by participants whereby they are allowed to revise items and receive feedback on the revisions until the items have no weaknesses. Although such an approach might be beneficial, it might not always be feasible because of the time and tutor workload constraints. A solution, however, might be in attracting more experienced item writers to act as mentors during the training and to provide individual participants with additional feedback. The positive role such mentorship might play in socialising novices into the item-writers' community is discussed in Section 5.4.2.

Participants universally requested to see more items of each type, both good examples and problematic ones. This study's findings suggest that a focus on examples might be a necessary feature of the early stage of item-writing skill acquisition and is a natural way in which a skill is acquired, as recognised in ACT-R. Anderson (1993b) warned, however, that such learning might sometimes result in "mistaken inferences" (p.88), which was the case for some participants in this study. For example, several participants assumed, by studying the writing prompt example, that the prompt instructions always had to contain two parts – an assumption which led participants to difficulties in producing their own writing prompts and resulted in lower item scores. One way of avoiding such misinterpretations is by providing multiple examples for each item type – something that is also advocated by Kim et al. (2010) who wrote that "item writers need … a range of sample items with different difficulty levels" (p.165). Having multiple examples might also help with highlighting the breadth of specification requirements, for instance by exemplifying each input text genre included in the specifications. However, tutors might need to be cautious not to make trainees over-reliant on examples at the expense of studying the specifications because no number of examples can cover all possible specifications' operationalisations, as well as no number of flawed items can account for all types of problems trainees might have with writing an item. Ultimately, the process of cognitive skill acquisition should result in a move away from example-based and towards rule-based performance (Anderson & Fincham, 1994), something that was also recognised by some participants in this study.

Finally, the online mode of training delivery deserves some discussion. This study revealed that trainees' digital literacy might affect course usefulness because how much a trainee is able to take from the course will be influenced by the trainee's ability to access materials, follow online tutorials, and collaborate with other participants virtually. One implication of this finding is that minimum digital literacy requirements might have to be set in order to participate in online training, as providing digital literacy training during an item-writing course might not always be feasible. Other ways to make training more accessible might be in using multiple formats of training materials (e.g. the same document as a .pdf, .doc, and as an online document) and of training input (e.g. as a *PowerPoint* with voice-over and as a text document). The training platform's user-friendliness should also be carefully considered.

5.3.5. Affordances and limitations of ACT-R for describing item-writing skills and their development

ACT-R theory of skill acquisition provided affordances for understanding the nature of item-writing skills and their development from cognitive perspective and as an individual process. The notions of declarative and procedural knowledge fundamental to ACT-R helped explain the nature of the knowledge that novice item writers have to acquire during the process of item-writing skill acquisition, while the model of skill development introduced in ACT-R served as a basis for explaining the process of item-writing skill development as it happened during the training course researched in this study; the model also allowed for speculations about further item-writing skill development that might happen after the training. However, to account for some of this study's findings, several other learning theories close to ACT-R had to be drawn on. In particular, although ACT-R acknowledges that cognitive skills are complex and are comprised of many components not developing simultaneously (Anderson, 1996), it does not offer an in-depth discussion of how components of a complex skill interact during the process of their acquisition. Speelman and Kirshner (2005), building on the concepts introduced in ACT-R, proposed a component model of skill acquisition that takes skill componentiality as its main premise. The model was helpful in discussing item-writing as a multi-component skill which components are acquired at different rates. Moreover, the notion of U-shaped skill development as a variation on the general process of skill acquisition, although compatible with the model of skill acquisition introduced in ACT-R, has not been afforded attention within this theory; therefore, to account for this study's findings on profile C participants, I drew on discussions of U-shaped development offered in Lesgold et al. (1988), McLaughlin (1990) and especially in Strauss and Stavy (1982).

At the same time, some other theories which had been considered (e.g. Skill Theory, Instructional-Design Theory, Situated Cognition), were not found useful for depicting this study's findings. For instance, Situated Cognition, which is often drawn on in relation to CoP, does not offer a concrete and comprehensive model for skills' development like ACT-R does – the Situated Cognition umbrella embraces a wide range of theories with varying views and approaches to learning. For example, the position of Situated Cognition is unclear with regard to transfer, with different definitions offered (see, e.g., Greeno et al., 1993; Young et al., 1997) and some researchers within the Situated Cognition umbrella claiming that transfer does not happen (Lave, 1988). In contrast, ACT-R theory offers a clear position on transfer of skills between tasks; the ACT-R view on skills transfer has been helpful in explaining this study's findings suggesting that some item-writing production rules were acquired faster than others due to positive transfer (see Section 5.3.3). Moreover, proponents of Situated Cognition believe that teaching abstractions in ineffective and, instead, advocate apprenticeship training as the only viable way of skill development (see, e.g., Collins et al., 1989). This position has attracted a lot of criticism. For example, Bereiter (1997) convincingly argued that accelerated pace of modern life calls for promoting learning that can be applicable in multiple situations, while the sole focus on situated apprenticeship-style training might lead to

> a future in which a small number of people have caught on to some secret of transferrable learning and thus are able to keep creating and adapting to new situations, while the rest of us find it increasingly difficult to cope (p.289).

Studies have demonstrated that abstract instruction has the ability to accelerate learning by provoking positive transfer (see, e.g., Biederman & Shiffrar, 1987), while a combination of abstract concepts with specific examples was found to be very effective way (see, e.g., Nesher & Sukenik, 1991; Reed & Bolstad, 1991). The training course in this study is an example of such a combination: as explained in Section 1.2.1, the scope of the training was defined based on the definition of LAL proposed by Fulcher (2012) and included both input on theoretical language testing principles and practical skills in item production.

This section discussed the cognitive dimension of item-writing skill acquisition with reference to the ACT-R theory. The next section focuses on the social dimension of item writing and item-writing training with reference to the CoP theory.

5.4 Learning to be an item writer as a social situated activity

5.4.1 Item writers as a community of practice (CoP)

Item writers have recently been recognised as a CoP (Constantinou et al., 2018; Ho, 2019) because involvement in the item-writing practice possesses all major CoP features. Item writing is a highly specialised and regularised domain where *shared understanding* (Wenger, 1998), one of the main CoP characteristics, is essential for item writers to do their work. "[P]assion for the domain" (Snyder & Wenger, 2010, p.110) is another important factor in being a CoP practitioner and, judging from my personal experience, one has to be passionate about item writing to sustain one's engagement with the community because item writing is a highly demanding activity that does not always bring a regular income. The practice of item writing, like in any CoP, involves a *shared repertoire* (Wenger, 1998) of frameworks, tools, documentation, and procedures: it is of crucial importance that each item writer adheres to the shared repertoire as the usability of the resulting items largely depends on this. Moreover, an item-writers' CoP normally enjoys a balance of *reification* (adhering to a strict set of rules and regulations) and *participation* (creating meaning by engaging in interactions), which Wenger (2012) called for; although item writers have to follow the guidelines, they also constantly engage in negotiations about them, which results in changes in the rules, procedures, and documentation. For example, as reported by Green & Hawkey (2011), writers of IELTS test items at Cambridge Assessment have to strictly adhere to item-writer guidelines, but the guidelines themselves "are periodically modified to reflect feedback from item writers as well as other stakeholders" (p.111).

Many testing bodies (e.g. Cambridge Assessment, ETS, The British Council, Trinity College London) employ item writers as freelancers who do their work from home and who do not often meet each other. However, it is not geographical proximity or socially visible boundaries but meaningful interaction that makes a group of practitioners a community (Wenger, 2000). Interaction and ensuing collaboration are inseparable parts of item-writing work to the extent that without them the item-writing activity cannot be carried out. The importance of collaboration in item writing was highlighted in the literature, for example Davidson and Lynch (2002) wrote that "the best tests … are the results of the collaborative effort of a group of people" (p.99), Green (2014) argued that "[t]he collective aspect [of item writing] is vital to successful outcomes" (p.43), while Ho's (2019) empirical results suggested that the ability to collaborate is necessary for item-writing.

Item-writing collaboration is essential for several reasons. Firstly, item writing is a distributed skill whereby no individual item writer has the competency to independently produce a set of test items but has to interact with reviewers and, possibly, item-writing colleagues to further craft their items - something that was highlighted in this study and in previous research. For example, Salisbury (2005) wrote that item writers "need to work as part of a complex domain system in order to bring their task to completion" (p.295). Secondly, item writers might have to collaborate with professionals from other fields, for example when producing items for LSP tests or when testing less-commonly taught languages (Ryan & Brunfaut, 2016). Moreover, the very process of item writing has an in-built collaborative element whereby items go through a review-revision cycle that involves multiple practitioners.

On-going learning is a necessary characteristic of a CoP (Wenger, 1998), and the fact that item-writer collaboration results in professional development was noted in the literature (Green & Hawkey, 2011; Ho, 2019). Traditionally, one would become an item writer while in employment by learning from more experienced practitioners (Ebel, 1963). Even though formal item-writing training is now recognised as essential, this study demonstrated that induction training alone might not be sufficient to make one an item writer because item-writing skills might require more practice to develop. This finding points to the importance of continuous development through engagement with the item-writing community during editing meetings, communication with item reviewers, peer-review, and collaborative item-writing sessions. Constantinou et al. (2018) wrote that "as a result of their socialisation in this community of practice, test writers appropriate the prevailing norms and discourse" (p.421). Ho (2019) found that "participation in the process of peer feedback and revision was a key aspect of item-writer development for the study participants" (p.65), with the participants themselves recognising the single importance of collaboration in learning to produce test items.

Green and Hawkey's (2011) study into the IELTS reading task production gives us an example of an item-writers' CoP in action. In the first stage, item writers select input texts and produce item drafts. Although the item writers work independently using item-writer guidelines, the guidelines are the result of on-going collaboration between production managers, item writers, and other stakeholders. In the second stage, a pre-editing panel made up of item-writer team leaders and production managers review the task drafts and return them to item writers with detailed guidance for revision: "pre-editing thus makes an important contribution to item writer training" (Green & Hawkey, 2011, p.112). Finally, the revised tasks are reviewed in an editing meeting which includes the item writers: "[t]hese meetings, and the opportunities they afford for interaction, further contribute to professional development" (Green & Hawkey, 2011, p.112).

5.4.2 Item-writing induction training as legitimate peripheral participation

Lave and Wenger (1991) emphasized that learning through *legitimate peripheral participation* in the CoP, understood as the process whereby "newcomers become part of a community of practice" (p.29), is superior to learning in the classroom. Therefore, although formal item-writing training is essential (Downing, 2006; Welch, 2006), for the training to be more effective it should incorporate some features of the actual item-writing practice as it happens within the CoP. Understanding of this seems to be on the rise in educational assessment; for example, the most recent study of the effectiveness of training medical faculty in writing MCQs (Gupta et al., 2020) concluded that short one-day sessions are not effective and recommended longitudinal intervention, hands-on exercises, one-to-one interaction, and engagement in the item-writing review-revision cycle as ways of improving the training's effectiveness. Notably, the suggestions aim to incorporate some characteristics of *legitimate peripheral participation* into formal item-writing training.

Admittedly, a formal item-writing training course can never be fully equal to *legitimate peripheral participation*, with differences between a learning community and a CoP widely discussed in the literature (see, e.g., Schwier & Daniel, 2008), the main one being the fact that a learning community is "an artificial construct created… with a didactic goal" (Bos-Ciussi et al., 2008, p.303). However, it is also acknowledged that, to maximise learning, a CoP can and must be cultivated within a learning community (Bos-Ciussi et al., 2008; Hibbert, 2008). It seems, then, that by integrating CoP activities into a training course one might create the necessary conditions for *legitimate peripheral participation* (*Figure 5-5)*. There are multiple ways of achieving this: by replicating the processes CoP members engage in; by using the communication channels characteristic of the CoP; by employing educators who are active CoP members; by involving other CoP members in the learning process; by using frameworks, routines, tools and documentation that are characteristic of the CoP. *Legitimate peripheral participation* can also be enhanced by avoiding the activities characteristic of formal education, for example lectures or comprehension check quizzes.

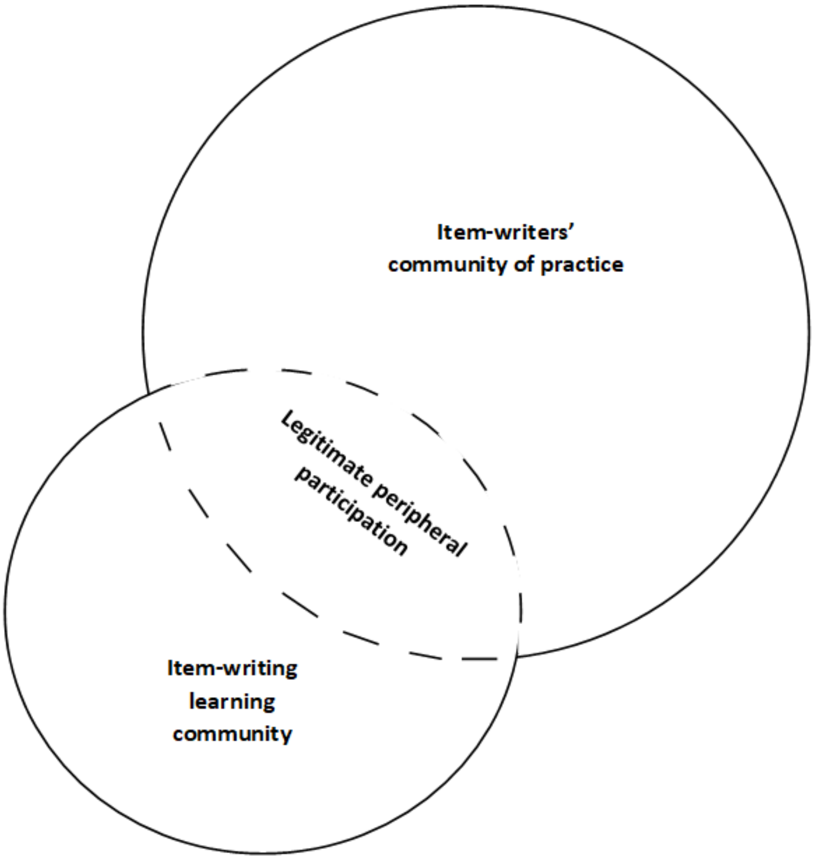

Item-writers' community of practice
Legitimate peripheral participation
Item-writing learning community

Figure 5-5. Legitimate peripheral participation in a formal item-writing training course

This study's findings indicate that participants appreciated the opportunities for *legitimate peripheral participation* they were given during the training. The item-writing practice was perceived as the most valuable part of the training. Participants particularly appreciated being provided with feedback after each item-writing event, not the least because they saw the review-revision cycle as a regular process of producing test items within the item-writers' CoP. Participants also valued multiple opportunities for collaboration during the training, and they perceived collaboration as an important feature of the item-writers' CoP. In their feedback, participants also offered ideas for increased collaboration, such as using Google docs or video-conferencing technology to produce items in pairs and groups.

The tutors, who were known to be practising item writers, were perceived as a bridge connecting the trainees with the wider item-writers' community. Therefore, participants particularly appreciated tutor feedback and tutor involvement in group discussions. They also looked for more ways of engaging with the tutors, for example through optional webinars, live Q&A sessions, or video tutorials whereby the tutors would relate their own personal experience of being item writers. Participants' desire to learn about the item-writers' CoP is also reflected in the fact that they wanted input on the item-writing market and job opportunities.

Participants' low appreciation of the few activities that originated from formal educational practices indicates that participants wanted the training course to be *legitimate peripheral participation* and not a formal educational event. For example, group discussion activities whereby participants were asked to read an article/chapter, answer comprehension questions, and discuss the answers in their group were not perceived as useful by several participants because these did not afford an opportunity for genuine communication through providing unique responses.

5.4.3 The role of induction training in socialising novices into the item-writers' CoP

The training course in this study offered participants multiple opportunities for *legitimate peripheral participation* in the item-writers' CoP, despite the course being a formal training event with a didactic goal. Firstly, the course aimed to replicate the processes item-writers' CoP members engage in while producing test items. As noted in the literature (e.g. Green & Hawkey, 2011), item writing normally involves both individual and group work; the present course's item-writing practice was organised so that first drafts of items were produced individually, the items were posted to the group for peer-feedback, participants then had a chance to revise the items before submitting them to the tutors who, acting as professional reviewers, provided detailed individual feedback. Finally, group feedback summaries were posted in the course space, which is similar to what is done within some item-writers' CoPs whereby team leaders or test managers compile lists of typical item-writing flaws and share them with the item writers. In their participant feedback, trainees expressed a clear preference for the combination of individual- and group-work, citing as one of the reasons that it is something that normally happens within an item-writers' CoP.

Collaboration among participants was a regular feature of each Module, whereby participants acted as item peer-reviewers and engaged in group activities. For example, in Module 1 they discussed item specifications and quality review checklists, while in Module 2 they collaboratively identified issues with grammar MC items. As revealed in the post-training interviews and participant feedback, group discussions were appreciated as a way of item-writer collaboration where one can share ideas, get a glimpse of others' item-writing approaches, and receive help and advice. Notably, 'high-achievers' reported taking advantage of such collaboration; prior to doing the post-training assignment, they revised their group discussions and the feedback they received from peers. One reason why 'high-achievers' were more successful than other participants might therefore be that they took full advantage of the collaboration opportunities the course provided.

However, participants also said that group-work was not always effective, as it depended heavily on the individual group members and general levels of activity within the group. Participants were allocated to a different group for each Module to ensure wider collaboration within the cohort and to avoid a situation whereby someone would stay in an inactive group throughout the whole course. However, joining a new group every two weeks required a period of familiarisation and social adjustment, which resulted in some participants feeling reluctant to provide negative feedback, as they noted in their post-training interviews. In their feedback, participants offered some suggestions for ensuring more active participation, for example by tutors encouraging participants to post; by rotating groups every other module instead of every module; by appointing group leaders who are given monitoring responsibilities; and by using more structured ways of communication, for example by allocating pairs within a group to provide each other with feedback on items. All these suggestions might be worth implementing in future item-writing training courses. However, there is also a risk that more tutor involvement and group regulation might result in tutor-domination (Hibbert, 2008) which might shift the training from *legitimate peripheral participation* to a formal educational activity.

Ensuring active participation might also be challenging because of the existence of so-called 'lurkers' – trainees who read online messages but do not take an active part in discussions. Several higher-education studies found that students who are not visibly engaged often still spend a significant amount of time on the course and engage in learning activities such as reading and thinking about other students' posts (Mazuro & Rao, 2011; Beaudoin, 2002). However, Beaudion (2002) also found that 'lurkers' generally have lower grades than their high-visibility peers. My personal observations during the course as the course tutor suggest that the majority of 'high-achievers' for each individual item type were also active peer-reviewers and group discussion participants in the Modules where the relevant item types were discussed. However, Ryan, who was also a 'high achiever', rarely participated in group discussions. This might suggest that active group participation might also depend on each trainee's personality and is not necessarily a pre-requisite for item-writing skill development. Ryan, for example, declined to be interviewed both before and after the training, which might point to his introverted nature. The latter did not prevent Ryan from developing his item-writing skills, however. Wenger (2002), in fact, advocated inviting "different levels of participation" (p.50) within a CoP, which might be interpreted as legitimizing lurker-style participation in training.

Item writers normally receive feedback on each item they are commissioned; regular feedback was also an important feature of the item-writing training in this study. Tutor feedback was perceived as most valuable for several reasons: it was very detailed, made use of quality-review checklists that are also a feature of professional item-writing practice, and was given by tutors who were also active item-writers' CoP members. In their course feedback, participants requested more tutor feedback via multiple cycles of review-and-revision. As discussed above (Section 5.3.4.2), course tutors might not be able to provide such additional feedback due to already big workloads. One solution, though, might be to involve experienced item writers other than tutors to act as mentors to course participants and to provide additional feedback. This might also enhance *legitimate peripheral participation,* as trainees will have a chance to work with more practising item writers who will offer different perspectives on the item-writing work. The positive role that such a combination of formal training and mentorship might play in developing item-writing skills was discussed in Smith and Geist (2020) who suggested the TERM model consisting of [T]raining, [E]valuation of past items, [R]ewriting past items based on the training input, and [M]entor feedback from the faculty experienced in item writing.

Besides tutor feedback, peer feedback was also appreciated by participants because it allowed them to see how other trainees approached writing items, but also because it was seen as a normal process within an item-writer community. Moreover, the participants appreciated the fact that peer-feedback was conducted in *Wechat* which allowed participants not only to submit/receive feedback but also to engage in discussions about it – something that was also reported as a feature of item editing meetings by Green and Hawkey (2011).

The socialisation of participants into the item-writers' CoP is evidenced through their acquisition of item-writing terminology. Before the training, participants either did not mention item-writing concepts such as 'construct' or replaced terminology with common lexis; for example, they used the word 'naturalness' for 'authenticity' or 'incorrect option' for 'distractor'. After the training, participants used item-writing terminology much more confidently. Among the terms the participants acquired are, for example, 'construct', 'grammar exponent', and 'distractor' when talking about grammar MC items (Section 4.3.2.1). The improved ability of the participants to talk about listening input text authenticity can serve as another example. While before the training participants rarely used the term 'authenticity' and had difficulty talking about spoken language features, after the training the participants confidently used the terms 'redundancy', 'ellipsis', 'false starts', 'fillers' and so on when talking about spoken language features of their listening texts (see Section 4.3.2.3).

The *legitimate peripheral participation* was also reinforced through the use of artefacts which, in my experience, are employed in professional item writing. For example, the item specifications, item templates, and quality-review sheets were modelled on real documents. The item-writing tools such as *Lextutor*, *Core Inventory*, and *Cohmetrix* are also used for actual item-writing at some exam boards. Finally, participants' high appraisal of the course's structure reflects Wenger's (2002) recommendation that, in order for a CoP to function, leaders should "create a rhythm for the community" (p.50). The training course in this study offered participants a predictable, though varied, pattern of activities that were repeated each Module and helped not only to create a rhythm for the training but also to replicate the predictable cycle of item-writing activities as they happen within an item-writers' CoP.

5.4.4. Affordances and limitations of CoP for describing item-writing skills and their development

CoP theory provided affordances for a discussion of item writing from a social situated perspective. The dimensions, elements and components of a CoP as explained in Wenger (1998) allowed for an in-depth understanding of the item-writers' CoP and its characteristics. The concept of legitimate peripheral participation (Lave & Wenger, 1991), central to the CoP theory, provided a model for discussing item-writing training researched in this study from a social perspective. Some other social learning theories which had been considered (e.g. Socio-cultural theory, Cultural-Historical Activity Theory), were not found as useful in discussing this study's findings. For instance, Cultural-Historic Activity Theory (Engestrom, 2014) previously used by Ngo (2016) to explore the factors that mediated the item-writing activity in his study (see Section 2.2.7) is well-suited to account for the role of tools and artifacts in complex human activities but does not provide adequate affordances to account for on-going learning as it happens within a community, something that is the strength of CoP with its concept of legitimate peripheral participation (Lave & Wenger, 1991).

At the same time, because the concept of legitimate peripheral participation was first introduced for apprenticeship training, it cannot be applied without modifications to formal learning such as the one researched in this study. This was acknowledged by researches who promote CoP principles within academia (see, e.g., Hoadley, 2012). Refinements to the CoP theory were proposed to account for learning in designed environments (e.g. Bos-Ciussi et al., 2008; Hibbert, 2008), and these additional sources were drawn on to complement the discussion in this chapter. Moreover, some phenomena specific to online learning, for example 'lurking', are not accounted for in the CoP theory, and studies that researched this phenomenon had to be additionally consulted (see Section 5.4.3 of this thesis).

Sections 5.3 and 5.4 discussed this study's findings with reference to two influential learning theories. The section that follows provides a discussion of a range of methodology-related issues. It offers reflections on this study's methodological decisions and how they improve on the methods used in some previous studies of item-writing training effectiveness.

5.5 The methodology of research into item-writing training effects

The overview of recent empirical research into item-writing training (see Literature Review Chapter, Section 2.2.7.1) concluded that the methodology for such research is still in its infancy. The present study's methodology might be viewed as a step forward towards establishing a valid methodology for research into item-writing training effects.

As discussed in Section 2.2.7.1, the studies into item-writing training effectiveness that evaluated items against a set of criteria considered only the total item evaluation scores and not scores on individual criteria (Dellinges & Curtis, 2017; Scott et al., 2019). Findings from the present study, however, demonstrated that total score statistics might be both uninformative and misleading. Wilcoxon signed-rank test results revealed significantly higher total scores for A2 and C1 grammar items and B1 listening tasks following the training. This finding could have led to sweeping claims of the training's effectiveness if the statistics for the scores on individual criteria had not been considered. The latter, however, demonstrated that improvement in the overall item quality was in a large part due to improvements in quality on the objectively-scored criteria, while the trainees' ability to comply with the subjectively-scored criteria was uneven following the training. On the other hand, the comparison of pre- and post-training total scores for B1 writing prompts produced no statistically significant results, which could have been interpreted as a training failure. However, a closer consideration of the descriptive statistics revealed that the pre-training scores were already high so there was less scope for changes, and for the differences in the scores pre- to post-training to be statistically significant.

This statistical test insensitivity was also discussed in Dellinges and Curtis (2017) who attributed it to the insensitivity of their two-band (yes/no) evaluation scale. They hypothesised that a wider band range "may increase the range of scores and provide higher sensitivity" (p.953). The present study, however, found that a three-band scale also, in many instances, resulted in very small differences between pre- and post-training item scores. Expanding the band range even further might prove problematic due to the need to produce multi-level band descriptors which, without extensive validation, might result in increased inconsistency of judgements.

In deciding on the evaluation scale design, it is also important to consider operational item-reviewing practices. Notably, Dellinges and Curtis (2017), Hamamoto Filho and Bicudo (2020), Naheem et al. (2012), Scott et al. (2019), and Tricio et al. (2018) all used two-band evaluation scales which were probably modelled on operational item reviewing where ‘conforms/does not conform’ (to the specifications) judgements are customarily made. The three-band scale used in this study, while widening the band range, does not deviate from item evaluation practices as it requires judges to make a decision to either accept an item as it is, to return it for revision, or to reject it – a decision-making process that is familiar to any item reviewer. Using a wider band range in item-writing training studies, however, might prove challenging for professional item reviewers who normally act as judges, because they would not have had experience in using such a scale in operational testing.

When devising item evaluation scales, the band range as well as the number of evaluation criteria should be carefully considered. Among the studies reviewed, only Naeem et al. (2012) developed a comprehensive 21-criterion MCQs evaluation scale, which was then adopted by Tricio et al. (2018). Dellinges and Curtis (2017), Hamamoto Filho and Bicudo (2020), and Scott et al. (2019) used 7-criterion scales which often conflated several different requirements into one criterion. The latter approach might lead to highly imprecise evaluations because the evaluation scale might fail to discriminate between items with many and with few flaws. Moreover, it might also be unclear to item writers what exactly the issue was with their item. As was found in the present study, few trainees were able to produce flawless items following the induction training, while many trainees produced better-quality items on many of the criteria, something that might have gone unnoticed if a less detailed evaluation scale had been employed. Adopting a suitably detailed wide-range evaluation scale is, therefore, of particular importance for studies that aim to investigate the effect of training on individual aspects of item quality.

Another important aspect of item evaluation methodology is the method for resolving judges’ disagreements. This, however, did not receive sufficient attention in the studies reviewed. Yurdakul et al. (2020) reported no method for resolving disagreements between the two judges in the study, while Hamamoto Filho and Bicudo (2020), Tricio et al. (2018), and Gupta et al. (2020) did not even report the number of judges the studies employed. Subjective judgements, however, often result in substantial disagreements (Bejar, 1983; O’Neill et al., 2019), something that was observed in both the present study and Dellinges and Curtis (2017) who reported a Kappa coefficient of 0.34 for the two judges. Dellinges and Curtis (2017) averaged the two raters’ evaluations, while in Scott et al. (2019) score discrepancies were adjudicated by a third rater. Whenever the latter method is used, the adjudicator’s superior professional qualifications have to be made clear (in operational testing it is normally a senior reviewer/reviewer trainer), something that was not reported in Scott et al. (2019). When the judges’ experience and qualifications are comparable, a better method of resolving disagreements might be the one used in the present study, that is using the medians of the judgements. As discussed in Section 3.5.1.4, the median is preferable over the mean as it helps avoid decimal points in item evaluations thus preserving the original scale whilst making score interpretations more meaningful. It should also be noted that, whenever possible, employing more than two judges is preferable as the score reliability increases together with the number of judges (Bejar, 1983). However, considerations of practicality typically prohibit using 20 judges as recommended by Bejar (1983), so a compromise has to be found between ensuring the reliability of judgements and keeping the study practicable.

Besides using a detailed evaluation scale and considering scores on individual evaluation criteria, the informativeness of the present study was increased through combining the results of several statistical measures. Apart from descriptive statistics and Wilcoxon signed-rank statistics, gain ratio statistics were also obtained (see Section 3.5.1.4). This proved particularly effective for detecting more nuanced changes in item quality from before to after the training. Moreover, gain ratio statistics provided information on the effect of the training on individual participants, which helped establish several trainee profiles (‘high-achievers’, ‘outliers’, and those whose post-training items scored lower than the pre-training ones). These important findings would not have been made had only the performance for the whole cohort been considered.

Finally, when researchers are interested in the effects of training on producing items of different types, or at different proficiency levels, it is advisable to include all item types / proficiency levels of interest in the study. The present study involved three item types - grammar MC items, writing prompts, and listening tasks - and found that higher post-training scores for one item type did not necessarily guarantee similarly high post-training scores for a different item type. Moreover, a trainee’s ability to produce grammar MC items also varied depending on the item’s target proficiency level (A2 or C1).

5.6 Summary

In this chapter, I interpreted the present study’s findings with reference to two theories – cognitive ACT-R theory of skill acquisition and social CoP theory of learning. From the cognitive perspective, I suggested that item writing is a multi-component skill with individual components acquired at different rates; in this study, the objective production rules developed in a more uniform and linear fashion across the whole cohort, while the formation of the subjective production rules occurred more slowly and was less uniform. The findings also suggest that there might exist production rules which are responsible for co-ordinating competing specification requirements, and that these rules might be the last to form.

This study's findings suggest that item-writing skills follow the process of acquisition similar for all complex cognitive skills: novice item writers first learn declarative item-writing information (such as item specifications, item-writing guidelines, example items). The declarative knowledge is then converted into procedural through item-writing practice. During knowledge compilation (or proceduralisation), item-writing production rules are formed. They then go through the process of tuning whereby useful rules are strengthened while less useful rules are rejected. It was further found that participants in this study followed the path of item-writing skill acquisition in different ways. Three participant profiles were discussed: (a) participants who experienced initial difficulties with interpreting declarative item-writing information but whose post-training items were of much higher quality; (b) 'high-achievers' who had better initial item-writing ability which resulted in fast declarative information interpretation and production rule formation; the improved quality of their post-training items suggested that the production rules were being successfully tuned; (c) participants who were initially as successful as 'high-achievers' but whose production rule tuning resulted in U-turns because these participants were more error-prone in their search for more effective item-writing approaches.

This study helped to confirm earlier suggestions that item writing is not a generic skill but is item-type and proficiency-level specific: all participants in this study displayed jagged profiles with, for example, one and the same person being an 'outlier', 'high-achiever', or following the U-shaped learning curve depending on the item type. Analysis of item-writing skill acquisition for individual item types demonstrated that writing MC grammar items might be susceptible to faster development, while producing listening tasks might take longer to learn. Producing writing prompts, though seemingly easy initially, might also take longer to perfect.

The study's findings demonstrated that induction item-writing training might benefit different trainees in different ways. Profile A participants benefitted from explicit instruction on using item specifications and item-writing tools. Profile B participants mostly benefitted from input on language assessment principles and suggestions on improving production rule effectiveness. The latter aspect of the course was also beneficial for profile C participants who, however, might need more item-writing practice and feedback for the learning to manifest in improved item quality (see Section 4.3.2).

A look at the social dimension of item writing revealed that item writers are a CoP whose practitioners have a shared understanding of their work, use a shared repertoire of artifacts, and are heavily dependent on collaboration. Because legitimate peripheral participation is seen as the best way of becoming a CoP practitioner, item-writing training courses might need to consider ways of incorporating features of legitimate peripheral participation into the training while also reducing the amount of activities typical of formal education. With regard to the training course in this study, it was found that such course features as extensive item-writing practice, regular group-work, the use of item-writing processes and documentation typical of operational item-writing, and employing tutors who are practicing item writers/reviewers helped maximise participants' chances for participation in the item-writers' community, as well as ensured participants' satisfaction with the training.

The present study might help advance the expert judgement research methodology used to investigate item-writing training effects. As the results revealed, statistical tests used to compare pre- and post-intervention scores might not be sufficiently sensitive in detecting nuanced changes in item quality. This insensitivity might be reduced by using a detailed evaluation scale comprised of a comprehensive set of criteria and a wider band range, although the latter should not be so wide as to increase the subjectivity of judgements. Moreover, scores on individual criteria – and not only total item scores – should be included in the analysis. Comparative statistical tests such as Wilcoxon signed-rank test should be supplemented with the analysis of descriptive statistics. Gain ratio statistics have also proved useful in determining the effect of training on individual participants. Finally, using human judgement inevitably leads to subjectivity, and ways to mitigate against such subjectivity should be carefully considered. Using many judges is not always practicable but, for the results to be valid, at least three judges are recommended. Their disagreements can be resolved either through adjudication, in which case the adjudicator's superior credentials should be clearly established, or by basing the analysis on the medians of individual judges' ratings.

Based on the findings and discussion provided in the last two chapters, the following chapter concludes this project by considering its implications, contributions, and limitations.

Chapter 6 Conclusion

6.1 Introduction

This Chapter concludes this thesis, which has investigated the item-writing skill development of twenty-five trainees as it happened during an existing item-writing training course. The study provided empirical evidence for the importance of training in developing the ability to produce language test items. The primary aim of this thesis was to gain insights into item-writing skills and their development through training. In particular, the study investigated how the quality of items produced by novice item writers changed from before to after the training. It also explored how participants approached item production before and after the training and investigated participants' perceptions of training usefulness in developing their item-writing skills.

This Chapter starts with a summary of the key findings for each of the research questions (Section 6.2). Then, the theoretical and methodological contributions of the research are outlined (Section 6.3). Next, implications for item-writing training are discussed (Section 6.4). Finally, limitations of the study are described (Section 6.5) and suggestions for further research are formulated (Section 6.6).

6.2 Summary of the main findings

This study empirically explored the development of item-writing skills as it happened during an online induction training course. Mixed methods were used to answer the three research questions of the study: statistical analyses of item evaluations, qualitative Grounded Theory analysis of participants' interviews, and statistical and thematic analyses of feedback questionnaire responses.

RQ1 – *How did the quality of items produced by novice item writers change from before to after an online item-writing training course?* – explored experts' judgements on the quality of three types of items produced by participants for pre- and post-training assignments. Descriptive statistics, Wilcoxon signed-rank tests, and gain ratio statistics were calculated to examine changes in item quality from before to after the training, as well as to explore individual item-writer variations. The findings from descriptive statistics revealed that participants already had some ability to produce test items prior to the training, but many participants produced better-quality items following the training. Wilcoxon signed-rank test results demonstrated that the total post-training scores for both A2 and C1 grammar items and for B1 listening tasks were statistically significantly higher compared to the pre-training ones. The improvement in the overall scores was in a large part due to an improvement in quality on objectively-scored criteria, while the changes in the scores on most subjectively-scored criteria were not statistically significant. No significant differences were found between the pre- and post-training scores on B2 writing prompts. Analysis of gain ratio statistics supported the observation that the improvement in quality on the objectively-scored criteria was greater and more uniform across the trainee cohort compared to the subjectively-scored ones. The analysis furthermore revealed four participant profiles with regard to changes in item quality on subjectively-scored criteria. Three of the profiles were further investigated in this thesis:

- *Profile A*: those whose item quality was low prior to the training but who produced better quality items following it;
- *Profile B*: those who produced good quality items before the training and whose post-training items were of even better quality;
- *Profile C*: those whose pre-training items were of reasonably good quality but whose post-training items scored one or several points lower.

The analysis also revealed that improvement in the quality on one item did not guarantee improvement in quality on other item types for the same participant, with most participants displaying jagged profiles.

To answer RQ2 – *How did the participants' item-writing skills develop through the training, as perceived by the participants in interviews?* – data from pre- and post-training interviews were analysed using a Grounded Theory approach. Participants reported item-writing as difficult on both occasions, with listening tasks characterised as the most and writing prompts as the least difficult. Prior to the training, participants mostly took guidance from example items, while after the training the specifications as a whole were their main point of reference. Awareness of objective requirements and the ability to use item-writing tools were generally sufficient in complying with the objective requirements. For subjective requirements, however, the analysis revealed different approaches to item writing by participants in different profile groups. Profile A participants demonstrated a much better understanding of specification requirements following the training, but their ability to produce items required further refinement. Profile B participants were generally able to understand the specifications pre-training, while the training improved their knowledge of language assessment principles underlying item production and helped them develop efficient item-writing approaches. Profile C participants followed a similar development path as profile B participants but encountered some problems with implementing the new knowledge following the training.

RQ3 – *What role did the participants perceive the training played in their item-writing skill development?* – explored qualitative and quantitative responses from four feedback questionnaires administered to participants at different times throughout the training. The analysis revealed participants' high overall satisfaction with the training, with the course structure receiving the highest praise. The course features which the participants found most useful included: input in language assessment principles, balance of theory and practice, variety of activities, extensive item-writing practice, and detailed feedback on items.

6.3 Contributions of this study

This study extends previous research in several ways. Three key theoretical contributions to the field are made: 1) providing insights into the nature of item-writing skills and introducing relevant terminology; 2) detailing the cognitive process of item-writing skill acquisition; and 3) providing insights into the socially-situated nature of item-writing skill development. This study also makes a methodological contribution to the field by advancing the research methodology to investigate item-writing training effects through expert judgements of item quality.

6.3.1 Theoretical contributions

6.3.1.1 Insights into the nature of item-writing skills

This study provides empirically-informed insights into the nature of item-writing skills. Findings from this study confirm previous suggestions in the educational measurement literature (Wesman, 1971) that there exists no unitary item-writing skill to produce items of all types. Instead, different items involve partly different skills from item writers (see Section 5.3.3). Furthermore, this study found that item-writing, similar to other complex cognitive skills (Speelman & Kirsner, 2005), is comprised of many components or production rules, using

ACT-R terminology. To my knowledge, this study is the first one to distinguish between two types of item-writing production rule: *objective production rules*, of which the execution has directly measurable outcomes, and *subjective production rules*, of which the execution relies on creative ability, general writing ability or knowledge of language assessment principles, and for which the outcomes require subjective judgement. Further, the study found that objective production rules can be formed fast and often simultaneously, while subjective production rules might take longer to form and are not formed at the same pace. It was further found that, during the item-writing process, several production rules are executed simultaneously and the quality of the resulting item largely depends on the item-writer's ability to co-ordinate the execution of production rules in one item, which might be a production rule in itself. Besides being of theoretical interest, these findings have direct implications for training item writers (see Section 6.4).

To enable the description of item-writing skills, I had to introduce and develop a set of terminology because, to my knowledge, no such terminology existed or was clearly defined prior to this study. In particular, I defined the term 'item-writing skills' and introduced the terms 'item-writing production rules', 'objective/subjective production rules', 'conflicting production rules', 'to co-ordinate/balance the execution of production rules'. The term 'production rule' was adopted from ACT-R (Anderson, 1993) while other terms were coined by me to reflect the findings from this study.

6.3.1.2 Insights into the cognitive dimension of item-writing skill development

This study has led to empirically-informed insights into the cognitive process of item-writing skill acquisition. Although an extensive body of research into the acquisition of various cognitive skills exists, to my knowledge, no research had been done prior to this study that looked into the acquisition of item-writing skills. This study found that the acquisition of item-writing skills happens according to the general principles of complex cognitive skill acquisition as described in the ACT-R theory (Anderson, 1993): the acquisition starts with interpreting declarative item-writing information (e.g., item-writing documentation, training input), the declarative knowledge is then converted into procedural through item-writing practice whereby production rules are formed, and the rules then undergo the process of tuning whereby weak production rules are discarded while good ones are strengthened and refined through trial-and-error exploration (see Section 5.3.2). The present study expanded on this general scheme by demonstrating that not all novices follow the path of item-writing skill acquisition in the same manner. In particular, it was found that the initial ability to interpret declarative item-writing information might differ among novices: some seem to be better able to understand and follow the requirements, while others might require explicit training. Moreover, it was found that production rule tuning does not happen in the same way for all novices: some seem to be successful at adopting more effective item-writing methods on first trial, while for others the process of trial-and-error exploration has to last longer.

Finally, this study also contributes to the understanding of how item-writing skills are acquired for different item types. It was found that grammar multiple-choice items might be susceptible to faster improvement while producing listening tasks might take longer to learn. Producing writing prompts, although seemingly easy initially, might take longer to perfect. These insights may serve to inform item-writing training (Section 5.3.4) and, consequently, improve the quality and the validity of resulting test items.

6.3.1.3 Insights into the social dimension of item-writing skill development

This study also further develops the understanding of item writing as a socially-situated activity through a Community-of-Practice lens. Following several recent studies that recognised that item writers are a CoP (Constantinou et al., 2018; Ho, 2019), the present study uses the CoP framework to explain the study's findings. This study's original contribution to the field lies in suggesting that induction item-writing training should offer opportunities for legitimate peripheral participation of novices in the item-writers' CoP. Furthermore, the findings lead to advice for relevant training activities, something that has implications for item-writing training design (see Section 6.4).

6.3.2 Methodological contributions

My review of previous research into item-writing training effects revealed that the relevant methodology is still in its infancy. This study has served to advance the methodology in several ways. Firstly, this study's results indicate that using statistical tests to compare pre- and post-training total item evaluation scores is not sufficiently informative as the tests might fail to detect more nuanced changes in item quality. To obtain more meaningful information about training effects, scores on individual evaluation criteria should be analysed using a combination of comparative and descriptive statistics. Secondly, this study's findings revealed a considerable variation in the item-writing skill development among the trainee cohort, something that is likely to be overlooked if only mean item evaluation scores are considered.

I therefore recommend employing statistical measures that take individual trainee variation into account. One such measure – gain ratio statistics – was successfully trialled in this study and can be recommended for future research.

This study employed an innovative research instrument – a detailed item evaluation scale that was specially designed for this study. The three-band scale is comprised of a comprehensive set of evaluation criteria for each item type. A comparison of the present study with previous studies that used an item evaluation scale (e.g. Dellinges & Curtis, 2017; Scott et al., 2019) revealed that the evaluation scale used in the present study produced more meaningful results.

Lastly, this study confirmed previous findings (e.g., Alderson & Lukmani, 1989; Bejar, 1983; O'Neill et al., 2019) that expert judgements are rarely in perfect agreement. Previous studies into item-writing training effects either avoided reporting such disagreements (e.g., Gupta et al., 2020; Yurdakul et al., 2020) or the methods used to resolve the disagreements were not methodologically sound (Dellinges & Curtis, 2017; Scott et al., 2019). The careful consideration given to judges' disagreements in this study resulted in a methodology that can be recommended for future research; the methodology involves deciding on the optimum number of judges for the study, selecting qualified judges, and creating a final dataset based on the medians of individual judges' ratings.

6.4 Implications of this study

There are several implications for the research carried out in this study, extending from the context of the present item-writing course to wider operational item writing, item-writer recruitment, and item-writing training.

The finding that there exists no unitary item-writing skill, but different items involve partly different skills from item writers (Section 5.3.3) may be used to inform operational item writing and item-writing training. For item writing, it cannot be assumed that an item writer who produces good quality items of one type will be equally good at producing items of a different type, which suggests the need for item-writer specialisation and/or targeted training. For item-writing training, this study's findings suggest that novices should be trained for a variety of item types, whereby not only their item-writing inclinations will be revealed but also chances for skills transfer maximised. This study also found that native or highly proficient users of a language cannot be assumed to have the knowledge of grammar or text genres, even if they are trained language teachers. As it is often impracticable to offer such broader linguistic training as part of an item-writing course, item-writer recruitment will want to ensure that applicants have such pre-existing knowledge and abilities. This could possibly be achieved by having detailed role specifications and being more stringent in evaluating applicants' suitability for the role.

This study's findings demonstrated that trainees differ in their initial item-writing ability, and the pace of their item-writing skills development varies. Therefore, for item-writing training to be effective, it should be designed to cater for diverse types of trainees. One way to ensure this is by including various types of input such as instruction in the use of item specifications and item-writing tools, language assessment principles, and effective item-writing approaches. Moreover, trainees in this study had different preferences as to how the input should be presented to them: many liked the information simplified as *PowerPoint* presentations, while some requested more in-depth academic reading. This suggests that item-writing training input should vary not only in terms of the content but also in the format of delivery.

This study also found that, while many trainees' item-writing production rules were being tuned following the training, for some trainees the production rules had only just started forming, which points to a large variation in training outcomes. No trainee in this study fully mastered item-writing skills following the training, which suggests that item-writing skill development is a lengthy process extending beyond the induction training. This finding might help in setting realistic expectations for induction item-writing training and raising the awareness that initial training has to be supplemented through on-going mentorship and further training whilst in employment. The finding also provides support for the suggestion that, in operational item writing, item review-and-revision cycles should always be implemented to ensure the quality of test items because, even if an organisation employs only trained and experienced item writers, they cannot be expected to always produce items ready for inclusion in a live test.

This study has helped to establish the view on the item-writing activity as a CoP, and to highlight the importance of collaboration in the item production process. Operational item writing should strive to maximise collaboration opportunities by encouraging item-writer collaboration during the writing process, peer-review, mentorship schemes, and experience-sharing events – something that was also requested by this study's participants (see Section

4.4.1). Furthermore, item-writing induction training should aim to maximize trainees' opportunities for legitimate peripheral participation in an item-writers' CoP whereby novices become part of the community from the start and, through participation in the community's activities, gradually move from its periphery towards the centre. This study has provided some practical suggestions on how this can be implemented, for example by replicating item-writing production processes during training, by employing tutors who are also practicing item writers, and by involving other experienced item writers as mentors for trainees.

Furthermore, this study has shed light on the role of specifications and example items in item-writing skill acquisition. The novice item writers in this study largely took guidance from examples, which seems to be a natural way of how a cognitive skill is acquired (Anderson, 1993). This suggests the need to provide novice item writers with a wide range of example items, both good and weak ones, which was also requested by participants in their feedback (Section 4.4.1). Moreover, to maximise examples' usefulness, they should reflect the breadth of the specification requirements; for instance, examples should be provided to illustrate each text genre required in the specifications. The process of item-writing skill acquisition, however, should result in a move away from example-based and towards specification-based performance. Therefore, item-writing training should highlight the importance of considering specification requirements and should help trainees in understanding specifications and using them effectively.

This study has also resulted in a range of more specific recommendations as to how induction item-writing training could be organised more optimally. For example, providing trainees with input in item-writing Dos and Don'ts does not seem to be sufficient – trainees should have plenty of item-writing practice followed by feedback, and the feedback should incorporate both positive and negative comments on the items (for more examples, see Discussion Chapter, Section 5.4.3).

Finally, the training in this study was conducted online over an extended period of time, which might be an optimal way of delivering such a training. Due to the need for extensive item-writing practice, as identified above, face-to-face workshops whereby trainees are gathered for one day or several consecutive days of input might not prove as effective. This study has resulted in a range of suggestions on how online item-writing training could be organised effectively. For example, it was found that trainees' digital literacy varies (see Section 4.4.1), which might affect training outcomes. This suggests the need to ensure, during the recruitment process, that all trainees are sufficiently digitally literate to follow the course. The training can also be made more accessible by using multiple formats of training materials and of training input to accommodate different participants' preferences and strengths. Moreover, the training platform should be given careful consideration: on one hand, it should be simple and easy to use, on the other hand, it should allow for various training modes including, for example, group work and peer review of items.

6.5 Limitations

Although this study sheds light on item-writing skill development through induction training, the study's limitations must be recognised to call for caution in generalising the results.

First, some general methodological limitations (also see Methodology Chapter) must be taken into account when interpreting the study's findings. The Pretest-Posttest study had a quasi-experimental design in the absence of a control group. Quasi-experimental designs may weaken studies' internal validity as it is difficult to claim with confidence whether the change occurred as a result of intervention of other variables such as incidental learning, natural subject maturation, or test practice effect (Baldwin & Berkeljon, 2010; Glass, 1965). Quasi-experimental designs, however, are commonly used in educational settings, and it was the only possible format in the context of this study due to ethical and practical considerations. As for alternative explanations of learning, incidental learning of item-writing is unlikely, while natural subject maturation and test practice effects cannot be ruled out but should not be considered a limitation because of this study's focus. Although this study included a course evaluation element (in particular, through the use of course feedback questionnaires), the study's aim was not to prove the particular effectiveness of the given training course but to explore item-writing skill development as it unfolds throughout induction training, be it as a result of instruction or because of other factors.

A further limitation concerns the study's participants. The sample of 25 participants resulted from practical considerations as it would be difficult to run a moderated online course with more than 25 trainees, which was already a very large group. Moreover, the 25 participants in this study were self-selected, with 35 people initially enrolled on the course but only 25 completing it. The participants who dropped out did so for personal reasons unconnected with the study, and they also left at an early stage. However, it is not possible to completely rule out the effect of self-selection on this study's results. To obtain more generalisable data, additional studies are needed with a diverse range of participants.

The present study was based on an existing item-writing training course and, therefore, is necessarily context-bound. Some of the results might have come from idiosyncratic characteristics of the study's participants as well as the course itself and may not be generalisable. Any study into item-writing training will, by necessity, have the same limitation. However, more generalisability could be achieved through conducting a range of similar studies in various training contexts and with a variety of trainees.

The low agreement found among the judges in this study is another limitation which suggests that caution is required in interpreting the findings. Low agreement is typical when subjective judgements are made (see e.g., Alderson & Lukmani, 1989; Bejar, 1983; O'Neill et al., 2019) and can be increased by employing a large number of judges (20 or more, according to Bejar, 1983), which was impossible due to this study's limited budget. In the present study, careful consideration was given to mitigating the effects of judges' disagreement by, firstly, selecting highly-proficient judges whose evaluations could be relied upon and, secondly, by creating a final dataset that served to maximise judges' agreement. I hope that subsequent studies can be carried out, especially studies commissioned by examination bodies with larger budgets, to help achieve better agreement through employing more judges.

The qualitative part of this study also has some limitations. Firstly, interviews were used to obtain information about participants' experiences of writing items for the two item-writing assignments. Interviewing is the most commonly used method of qualitative data collection in item-writing studies. However, it has its limitations: because some time had passed between the item-writing event and the interview, the interviewees might have had more limited recollections of their actions, while the recollections they had might not have been fully accurate. An alternative method of data collection would be think-aloud which, however, has its own disadvantages, the main one being changing the nature of the process under investigation. I decided, therefore, to use the retrospective interview method as less intrusive. However, further research which would combine think-aloud and interview methodology might be of benefit: it would allow for the methods' comparison to determine which method of data collection is more suitable in studies investigating item-writing processes.

The qualitative findings in this study come from the information provided by participants in the interviews. However, the data is based on what the participants chose to talk about. The fact that participants chose not to talk about a particular aspect of item-writing does not necessarily mean they were not aware of it or did not attend to it. This study's data triangulation aimed to mitigate this limitation by considering both item evaluations and interviews when deciding on a participant’s trajectory of item-writing skill development. This combination proved highly beneficial. For example, the interviews revealed that the participants whose items scored lower following the training did, however, make progress in their item-writing skill acquisition (see Section 4.3.2) – a finding that would have been overlooked had only item evaluations been considered. Another example is that the post-training items were of much higher quality on the objective criteria, while the participants offered limited discussions of objective requirements in their interviews. Had only the interviews been considered, the participants’ progress in complying with the objective requirements might not have been detected. I acknowledge, however, that there might have been some aspects of item writing that participants did not explicitly mention and that might have escaped this study’s attention.

A final point I would like to mention is the multiple roles I took in this study. Besides being the researcher, I was also the course designer and a tutor. This study, however, was not contract research but evolved from my personal interest in the topic and was done by me as a full-time self-funding doctoral student, so there was no pressure to produce proof of course effectiveness. Even so, an unconscious bias cannot be wholly ruled out. My other role in this study was as one of the expert judges. I took this role when it became evident that conflict resolution of two judges’ ratings was needed while recruiting an additional judge was impossible due to the lack of funding. To make sure that I had no memory of the items, I allowed a half-year gap between the training and the time I rated the items, which were also randomised and anonymised.

Despite these limitations, the findings of this study have contributed to a deeper understanding of item-writing skill development through induction training.

6.6 Suggestions for future research

Informed by the findings and limitations outlined above, several suggestions for further research can be formulated. This study’s aim was to investigate the nature of item-writing skills and their development through an online induction item-writing training course. Subsequent studies should be carried out to hopefully confirm the findings and to further investigate the cognitive and social dimensions of item-writing skill development. It would be good if a larger body of data was collected via replicating this study in various item-writing contexts, for a variety of item types, and with a variety of trainees. Moreover, to complement the findings from this study, other research methods could be employed. For example, think-aloud protocols could be collected from participants or video observations of item-writing sessions could be conducted.

The training course itself deserves more research attention than was feasible in this study. The course can be further researched through: 1) analysis of training materials and documentation; 2) analysis of the items produced by participants during the training and not only before and after it; and 3) analysis of participants' interactions during the training, in particular, online group discussions and peer-feedback.

This study was necessarily limited to investigating the item-writing skill development as it happened during the training, which left any skill development that might have occurred after the training unexplored. Longitudinal studies are needed to investigate how item-writing skills develop following induction training. It seems particularly interesting to investigate how much time is needed for novice item writers to become experts, and whether all trainees who are initially successful eventually achieve expert status.

One of this study's findings was variations in the trajectory of item-writing skill development for different trainees. Further research is needed to investigate the sources of such variation. For example, this study's tentative explanation for the lower scores received by some participants post-training was their excessive focus on a particular item aspect combined with lower working-memory capacity that limited the participants' ability to attend to other aspects of the item. This is only speculative, though, because this study did not measure the working memory capacity of the participants. Research is needed that investigates individual trainee characteristics and relates the findings to the development of the trainees' item-writing skills.

More research is also needed to explore the social dimension of item writing by investigating the effectiveness of various types of collaboration that happen within operational item-writing settings. Research into the benefits of the apprenticeship model of item-writing training would also be valuable as it would allow the determining of what training formats are more beneficial for item-writing skill development.

Finally, the training course offered to participants in this study was necessarily limited by the current state of knowledge about item writing. Although I made every attempt to find out about most beneficial approaches to item production, only limited knowledge exists, or if it exists, it is not documented or publicly available. As noted by Anderson and Corbett (1993), any complex cognitive skill training should be preceded with a careful investigation into the components of the said skill. Unfortunately, such investigations of item-writing skills are lacking at the moment. Therefore, to maximise item-writing training effectiveness, research into item-writing skills of expert item writers is urgently needed. Such research should aim to provide a detailed account of how item writing occurs for items of different types, and to document more effective item-writing approaches which can then be focused on within the training.

References

Abdulghani, H. M., Ahmad, F., Irshad, M., Salah Khalil, M., Al-Shaikh, G. K., Syed, S., & Haque, S. (2015). Faculty development programs improve the quality of Multiple-Choice Questions items' writing. *Scientific Reports*, *5*, 1–7. DOI: 10.1038/srep09556

Alderson, J. C. (1993). Judgements in language testing. In D. Douglas and C. Chapelle (Eds.), *A new decade of language testing research: Selected papers from the 1990 language testing research colloquium* (pp.46-57). TESOL.

Alderson, J. C. (2010). A survey of aviation English tests. *Language Testing, 27*(1), 51-72. DOI: 10.1177/0265532209347196

Alderson, J. C. (2010). *Assessing Reading.* Cambridge University Press.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.

Alderson, J. C., & Kremmel, B. (2013). Re-examining the content validation of a grammar test: The (im)possibility of distinguishing vocabulary and structural knowledge. *Language Testing, 30*(4), 535-556. DOI: 10.1177/0265532213489568

Alderson, J. C., & Lukmani, Y. (1989). Cognition and reading: Cognitive levels as embodied in test questions. *Journal of reading in a foreign language, 5,* 253-270.

Al-Lawati, Z. (2014). *An investigation of the characteristics of language test specifications and item writer guidelines, and their effect on item development.* Unpublished doctoral thesis, Lancaster University, UK.

Ambridge, B., & Lieven, E. V. M. (Eds.) (2011). *Child language acquisition: Contrasting theoretical approaches.* Cambridge University Press.

American Psychological Association. (n.d.). Capability. In *APA dictionary of psychology*. Retrieved August 21, 2020 from https://dictionary.apa.org/ capability.

American Psychological Association. (n.d.). Cognitive ability. In *APA dictionary of psychology*. Retrieved August 21, 2020 from https://dictionary.apa.org/cognitive-ability.

American Psychological Association. (n.d.). Skill. In *APA dictionary of psychology*. Retrieved August 21, 2020 from https://dictionary.apa.org/skill.

Anderson, J. R. (1987). Skill acquisition: Compilation of weak-method problem solutions. *Psychological Review*, *94*(2), 192–210.

Anderson, J. R. (1993a). Knowledge representation. In J. R. Anderson (Ed.), *Rules of the mind* (pp. 17–44). Lawrence Erlbaum Associates.

Anderson, J. R. (1993b). Learning. In J. R. Anderson (Ed.), *Rules of the mind* (pp. 69–92). Lawrence Erlbaum Associates.

Anderson, J. R. (1993c). Performance. In J. R. Anderson (Ed.), *Rules of the mind* (pp. 45–68). Lawrence Erlbaum Associates.

Anderson, J. R. (1993d). Production systems and the ACT-R theory. In J. R. Anderson (Ed.), *Rules of the mind* (pp. 1-17). Lawrence Erlbaum Associates.

Anderson, J. D. (1996). *The architecture of cognition*. Laurence Erlbaum Associates.

Anderson, J. D. (2010). *Cognitive psychology and its implications* (7th ed.). Worth Publishers.

Anderson, J. R. (Ed.) (1981). *Cognitive skills and their acquisition*. Lawrence Erlbaum Associates.

Anderson, J. R. (Ed.) (1993). *Rules of the mind.* Lawrence Erlbaum Associates.

Anderson J. R., Belezza, F. S., & Boyle, C. F. (1993a). The geometry tutor and skill acquisition. In J. R. Anderson (Ed.), *Rules of the mind* (pp.165-182). Lawrence Erlbaum Associates.

Anderson, J. R., Conrad, F. G., Corbett, A. T., Fincham, J. M., Hoffman, D., & Wu, Q. (1993b). Computer programming and transfer. In J. R. Anderson (Ed.), *Rules of the mind* (pp.205-234). Lawrence Erlbaum Associates.

Anderson, J. R., & Corbett, A. T. (1993). Tutoring of cognitive skill. In J. R. Anderson (Ed.), *Rules of the mind* (pp. 235-278). Lawrence Erlbaum Associates.

Anderson, J. R., & Fincham, J. M. (1994). Acquisition of procedural skills from examples. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(6), 1322–1340. DOI: 10.1037/0278-7393.20.6.1322

Anderson, J. R., & Singley, M. K. (1993). The identical elements theory of transfer. In J. R. Anderson (Ed.), *Rules of the mind* (pp. 183-204). Lawrence Erlbaum Associates.

Aptis (2015). *Aptis General technical manual version 1.0* (Technical Report). Retrieved August 15, 2019 from https://www.britishcouncil.org/aptis-general-technical-manual-version-10.

Argyris, C., & Schon, D.A. (1978). *Organizational learning: A theory of action perspective.* Addison-Wesley.

Attwell, G. (2006). *Evaluating E-Learning: A guide to the evaluation of e-learning*. Evaluate Europe Handbook Series, Vol. 2, European Commission. Retrieved September 10, 2019 from http://pontydysgu.org/wp-content/uploads/2007/11/eva_europe_vol2_prefinal.pdf.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing, 19*(4), 453-476. DOI: 10.1191/0265532202lt240oa

Bachman, L. F., Davidson, F., & Milanovic, M. (1996). The use of test method characteristics in the content analysis and design of EFL proficiency tests. *Language Testing, 13*, 125-150.

Bachman, L. & Palmer, A. (2012). *Language assessment in practice*. Oxford University Press.

Bainbridge Frymier, A., Shulman, G.M., & Houser, M. (1996). The development of a learner empowerment measure. *Communication Education, 45*(3), 181-199. DOI:10.1080/03634529609379048

Baldwin, S. & Berkeljon, A. (2012). Quasi-experimental design. In N. J. Salkind (Ed.), *The encyclopaedia of research design* (pp.1171-1176). SAGE Publications.

Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. *Language Testing, 28 (1),* 51–75. DOI: 10.1177/ 0265532210376379

Beaudoin, M. F. (2002). Learning or lurking? Tracking the "invisible" online student. *The Internet and Higher Education, 5*(2), 147-155. DOI: 10.1016/S1096-7516(02)00086-6

Bejar, I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement, 7*, 303–310.

Bereiter, C. (1997). Situated cognition and how to overcome it. In D. Kirshner and J.A. Whitson (Eds.), *Situated cognition: Social, semiotic, and psychological perspectives* (pp.281-300). Erlbaum.

Biederman, I., & Shiffrar, M. (1987). Sexing day-old chicks. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*(4), 640-645.

Blackmore, C. (2010). Managing systemic change: Future roles for social learning systems and Communities of Practice? In C. Blackmore (Ed.), *Social learning systems and communities of practice* (Vol. 12, pp. 201–218). The Open University. DOI: 10.1007/978-1-84996-133-2_12

Blessing, S. B., & Anderson, J. R. (1996). How people learn to skip steps. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(3), 576-598. DOI: 10.1037/0278-7393.22.3.576

Block, D. (1998). Tale of a language learner. *Language Teaching Research, 2*(2), 148-176.

Boeije. H. (2010). *Analysis in qualitative research*. SAGE Publications.

Bos-Ciussi, M., Augier, M., & Rosner, G. (2008). Learning communities are not mushrooms – or – How to cultivate learning communities in higher education. In C. Kimble, P. Hildreth, and I. Bourdon (Eds.), *Communities of Practice: Creating learning environments for educators* (Vol.2, pp.287-308).

Brown, D. (2010). *Language assessment principles and classroom practices*. Pearson Education.

Bryan, W. L., & Harter, N. (1899). Studies on the telegraphic language: The acquisition of a hierarchy of habits. *Psychological Review, 6*, 345-375.

Buck, G. (2001). *Assessing listening*. Cambridge University Press.

Buck, G. (2009). Challenges and constraints in language test development. In J. C. Alderson (Ed.), *The politics of language education: Individuals and institutions* (pp.166-184). Multilingual Matters.

Budiu R., & Anderson J. R. (2002). Comprehending anaphoric metaphors. *Memory & Cognition, 30*, 158–165. DOI: 10.3758/BF03195275

Byrne, M. D., & Kirlik, A. (2005). Using computational cognitive modeling to diagnose possible sources of aviation error. *International Journal of Aviation Psychology, 15*, 135–155. DOI: 10.1207/s15327108ijap1502_2

Cambridge English Language Assessment: Producing exams. retrieved February 17, 2017 from http://www.cambridgeenglish.org/why-cambridge-english/producing-exams/.

Carter, R. A., & McCarthy, M. J. (2007). *Cambridge grammar of English: Spoken and written English grammar and usage.* Cambridge University Press.

Chapman, D., & Stone, S. J. (2010). Measurement of outcomes in virtual environments. *Advances in Developing Human Resources, 12*(6), 665-680. DOI: 10.1177/1523422310394792

Charmaz, K., & Bryant, A. (2011). Constructing grounded theory analyses. In D. Silverman (Ed.). *Qualitative research*, 4th ed. (pp. 347-362). SAGE Publications.

Chi, M. T. H., Glaser, R., & Farr, M. J. (Eds.). (1988). *The nature of expertise.* Lawrence Erlbaum Associates.

Chow, S. (2010). Experimental design. In N. J. Salkind (Ed.), *The encyclopaedia of research design* (pp.447-452). SAGE Publications.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardised assessment instruments in psychology. *Psychological Assessment, 6*(4), 284-290.

Cohen, L., Manion, L., & Morrison, K. (2011). *Research methods in education* (7th ed.). Routledge.

Collins, A., Brown, J.S., & Newman, S. (1989). Cognitive apprenticeship: Teaching students the craft of reading, writing, and mathematics. In L.B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honour of Robert Glaser* (pp.453-494). Erlbaum.

Constantinou, F., Crisp, V., & Johnson, M. (2018). Multiple voices in tests: Towards a macro theory of test writing. *Cambridge Journal of Education*, *48*(4), 411–426. DOI: 10.1080/0305764X.2017.1337723

Corbin, J., & Strauss, A. L. (2015). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (4th ed.). SAGE Publications.

Creswell, J. (2013). *Qualitative inquiry & research design: Choosing among five approaches* (3rd ed.). SAGE Publications.

Creswell, J. (2014). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Pearson.

Creswell, J., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed). SAGE Publications.

Cronbach, L. J. (1970). Book review. *Psychometrika, 35,* 509–511.

Davidson, F. (2012). Test specifications and criterion referenced assessment. In G. Fulcher and F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 197-207). Routledge. DOI: 10.4324/9780203181287.ch13

Davidson, F., & Fulcher, G. (2007). The Common European Framework of Reference (CEFR) and the design of language tests: A matter of effect. *Language Teaching, 40*, 231-241. DOI:10.1017/S0261444807004351

Davidson, F., & Lynch, B. K. (2002). *Testcraft: A teacher's guide to writing and using language test specifications*. Yale University Press.

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

de Jong, J. (2008, August). Procedures for training item writers and human raters. Paper presented at the *EALTA Annual Conference*, Athens, Greece.

DeKeyser, R. M. (2007). Skill acquisition theory. In B. VanPatten and J. Williams (Eds.), *Theories in second language acquisition: An Introduction* (pp. 97–113). Routledge.

DeKeyser, R. M. (2009). Cognitive-psychological processes in second language learning. In M.H. Long and C.J. Doughty (Eds.), *The handbook of language teaching* (pp.119-138). Blackwell Publishing Ltd. DOI: 10.1002/9781444315783.ch8

Dellinges, M. A., & Curtis, D. A. (2017). Will a short training session improve multiple-choice item-writing quality by dental school faculty? A pilot study. *Journal of Dental Education*, *81*(8), 948–955. DOI: 10.21815/JDE.017.047

Dornyei, Z. (1990). Conceptualizing motivation in foreign-language learning. *Language Learning, 40*(1), 45-78. DOI:10.1111/j.1467-1770.1990.tb00954.x

Downing, S. M. (2006). Twelve steps for effective test development. In T. M. Haladyna and S. M. Downing (Eds.), *Handbook of test development* (pp.3-25). Lawrence Erlbaum Associates.

Downing, S. M., & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education, 10*(1), 61-82. DOI: 10.1207/s15324818ame1001_4

EALTA guidelines for good practice in language testing and assessment (2006). Retrieved November 15, 2017 from http://www.ealta.eu.org/documents/archive /guidelines /English.pdf.

Ebel, R. L. (1951). Writing the test item. In E. Lindquist (Ed.), *Educational measurement* (pp.185-249). American Council on Education.

Edwards, A. (2005). Let's go beyond community and practice: The many meanings of learning by participating. *The Curriculum Journal, 16*(1), 49-65.

Elliott, M., & Wilson, J. (2013). Context validity. In A. Geranpayeh and L. B. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening* (pp.152-241). Cambridge University Press.

Engestrom, Y. (2014). *An activity-theoretical approach to developmental research.* Cambridge University Press.

Ericsson, K.A., & Charness, N. (1997). Cognitive and developmental factors in expert performance. In P. J. Feltovich, K. M. Ford, and R. R. Hoffman (Eds.), *Expertise in context: Human and machine* (p. 3–41). American Association for Artificial Intelligence; The MIT Press.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). SAGE Publications.

Field, J. (2013). Cognitive validity. In A. Geranpayeh and L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening* (pp.77–151). Cambridge University Press.

Field, J. (2019). *Rethinking the second language listening test: From theory to practice.* Equinox.

Fischer, K. W. (1980). A theory of cognitive development: The control and construction of hierarchies of skills. *Psychological Review, 87*(6), 477–531. DOI:10.1037/0033-295X.87.6.477

Fitts, P. M., & Posner, M. I. (1967). *Human performance*. Brooks/Cole Publishing Company.

Flower, L., & Hayes, J.R. (1981). A cognitive process theory of writing. *College Composition and Communication, 32*(4), 365-387.

Fortus, R., Corriat, R., & Fund, S. (1998). Prediction of item difficulty in the English subtest of Israel's inter-university psychometric entrance test. In A. Kunnan (Ed.), *Validation in Language Assessment* (pp.61-87). Erlbaum.

Fu, W-T, & Pirolli, P. (2007). SNIF-ACT: A cognitive model of user navigation on the World Wide Web. *Human-Computer Interaction*, *22*(4), 355–412. DOI: 10.1080 /07370020701638806

Fulcher, G. (1997). Text difficulty and accessibility: Reading formulae and expert judgement. *System, 25*(4), 497-513.

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly, 9*(2), 113-132. DOI: 10.1080/15434303.2011.642041

Fulcher, G., & Davidson, F. (2009). Test architecture, test retrofit. *Language Testing, 26*(1), 123-144. DOI: 10.1177/0265532208097339

Fulkerson, D., Mittelholtz, D. J. & Nichols, P. D. (2009, April). The psychology of writing items: Improving figural response item writing. Paper presented at *the Annual Meeting of the American Educational Research Association,* San Diego, California.

Fulkerson, D., & Nichols, P. (2010). Capturing Expert Item Writers' Item Writing Expertise. *Test, Measurement & Research Services Bulletin*, 15, 1-3.

Fulkerson, D., Nichols, P., & Mittleholtz, D. (2010, May). What item writers think when writing items: Towards a theory of item writing expertise. Paper presented at *the Annual Meeting of the American Educational Research Association*, Denver, Colorado.

Fulkerson, D., Nichols, P., & Snow, E. (2011, April). Expanding the model of item-writing expertise: Cognitive processes and requisite knowledge structures. Paper presented at *the Annual Meeting of the American Educational Research Association*, New Orleans, Louisiana.

George, J. W., & Cowan, G. (1999). *A handbook of techniques for formative evaluation.* Kogan Page.

Geranpayeh, A., & Taylor, L. B. (Eds.) (2013). *Examining listening: Research and practice in assessing second language listening*. Cambridge University Press.

Gilmore, A. (2004). A comparison of textbook and authentic interactions. *ELT Journal, 58*(4), 363-374. DOI: 10.1093/elt/58.4.363

Gilmore, A. (2015). Research into practice: The influence of discourse studies on language descriptions and task design in published ELT materials. *Language Teaching, 48*(4), 506-530. DOI: 10.1017/S0261444815000269

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research.* Aldine.

Glaser, R., & Chi, M. T. H. (1988). Overview. In M. T. H. Chi, R. Glaser and M. J. Farr (Eds.), *The nature of expertise* (pp. xv–xxvii). Lawrence Erlbaum Associates.

Glaser, R., Chi. M.T.H., & Farr, M.J. (1988). *The nature of expertise.* Lawrence Erlbaum Associates.

Glass, G. V. (1965). Evaluating testing, maturation, and treatment effects in a pretest-posttest quasi-experimental design. *American Educational Research Journal, 2*(2), pp.83-87.

Green, A. (1998). *Verbal protocol analysis in language testing research*. Cambridge University Press.

Green, A. (2014). *Exploring language assessment and testing: Language in action*. Routledge.

Green, A., & Hawkey, R. (2011). Re-fitting for a different purpose: A case study of item writer practices in adapting source texts for a test of academic reading. *Language Testing, 29*(1), 109-129. DOI: 10.1177/0265532211413445

Green, A., & Hawkey, R. (2012). An empirical investigation of the process of writing Academic reading test items for the International English Language Testing System. In L. Taylor

& C. Weir (Eds.), *IELTS collected papers 2: Research in reading and listening assessment* (pp.270-378). Cambridge University Press

Green, A., & Jay, D. (2005). Quality assurance and quality control: Reviewing and pretesting examination material at Cambridge ESOL. *Research Notes*, *21*, 5–7.

Green, J., Caracelli, V., & Graham, W. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis, 11*(3), 255-274.

Green, R. (2017). *Designing listening tests*. Palgrave Macmillan.

Greeno, J.G., Moore, J.L., & Smith, D.R. (1993). Transfer of situated learning. In D.K. Detterman and R.J. Sternberg (Eds.), *Transfer on trial: Intelligence, cognition, and instruction* (pp.99-167). Ablex Publishing.

Gupta, P., Meena, P., Khan, A. M., Malhotra, R. K., & Singh, T. (2020). Effect of faculty training on quality of multiple-choice questions. *International Journal of Applied and Basic Medical Research*, *10*, 210–214. DOI: 10.4103/ijabmr.IJABMR_30_20

Haladyna, T. M. (2006). Roles and importance of validity studies in test development. In S. M. Downing and T. M. Haladyna, T. M. (Eds.), *Handbook of test development* (pp.739-758). Lawrence Erlbaum Associates.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17–27.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items.* Routledge.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309-333. DOI: 10.1207/S15324818AME1503_5

Hamamoto Filho, P. T., & Bicudo, A. M. (2020). Improvement of faculty's skills on the creation of items for progress testing through feedback to item writers: A successful experience. *Revista Brasileira De Educação Médica*, *44*(1), 486. DOI: 10.1590/1981-5271v44.1-20190130.ing

Hambleton, R., & Jirka, S. (2006). Anchor-based methods for judgmentally estimating item statistics. In S. M. Downing and T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 399–420). Erlbaum.

Hamp-Lyons, L., & Mathias, S. P. (1994). Examining expert judgements of task difficulty on essay tests. *Journal of Second Language Writing, 3*(1), 49-68.

Hayes-Roth, F., Klahr, P., & Mostow, D. J. (1981). Advice taking and knowledge refinement: An iterative view of skill acquisition. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 231–254). Lawrence Erlbaum Associates.

Heaton, J. B. (1990). *Writing English language tests*. Longman.

Hibbert, K. (2008). Virtual Communities of Practice: A vehicle for meaningful professional development. In C. Kimble, P. Hildreth, and I. Bourdon (Eds.), *Communities of Practice: Creating learning environments for educators* (Vol.2, pp.127-148).

Ho, E. C. P. (2019). *An exploration of the LAL development of pre-service ESL teachers through the processes of item writing*. Unpublished master's dissertation, University of Illinois at Urbana-Champaign, Illinois.

Hoadley, C. (2012). What is a Community of Practice and how can we support it? In S. Land and D. Jonassen (Eds.), *Theoretical foundations of learning environments* (pp.286-299). Taylor & Francis Group.

Hughes, A. (2003). *Testing for language teachers*. Cambridge University Press.

Hughes, J., Jewson, N., & Unwin, L. (2007). Introduction. Communities of practice: A contested concept in flux. In J. Hughes, N. Jewson, and L. Unwin (Eds.), *Communities of practice: Critical perspectives* (pp.1-16). Routledge.

IELTS (2007). *The IELTS Question Paper Production Process*, retrieved April 15, 2016 from http://www.ielts.org.

Ingham, K. (2008). The Cambridge ESOL approach to item writer training: The case of ICFE listening. *Research Notes, 32*, 5-9.

International Language Testing Association (ILTA) guidelines for practice (2007). Retrieved November 15, 2017 from http://www.iltaonline.com/images/pdfs/ILTA_Guidelines.pdf .

Iramaneerat, C. (2012). The impact of item writer training on item statistics of multiple-choice items for medical student examination. *Sirirai Medical Journal*, *64*(6), 178–182.

Johnson, J. M. (2002). In-depth interviewing. In J. Gubrium & Holstein, J.A. (Eds.), *Handbook of interview research* (pp.104-119). SAGE Publications.

Johnson, M., Constantinou, F., & Crisp, V. (2017). How do question writers compose external examination questions? Question writing as a socio-cognitive process. *British Educational Research Journal*, *43*(4), 700–719. DOI: 10.1002/berj.3281

Jozefowicz, R., Koeppen, B., Case, S., Galbraith, R., Swanson, D., & Glew, R. (2002). The quality of in-house medical school examinations. *Academic Medicine: Journal of the Association of American Medical Colleges, 77*(2), 156-61.

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp.17-64). American Council on Education.

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

Kim, J., Chi, Y., Huensch, A., Jun, H., Li, H., & Roullion, V. (2010). A Case Study on an Item Writing Process: Use of Test Specifications, Nature of Group Dynamics, and Individual Item Writers' Characteristics. *Language Assessment Quarterly*, *7*(2), 160–174. DOI: 10.1080/15434300903473989

Kimble, c., Hildreth, P., & Bourdon, I. (Eds.) (2012). *Communities of Practice: Creating learning environments for educators*. Information Age Publishing, Inc.

Kirkpatrick, D. J., & Kirkpatrick, J. D. (2006). *Evaluating training programmes.* Berrett-Koehler Publishers, Inc.

Kirwan, C. (2013). *Making sense of organisational learning: Putting theory into practice.* Taylor & Frances Group.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*, 155-163, DOI: 10.1016/j.jcm.2016.02.012

Krapp, A. (1999). Interest, motivation and learning: An educational-psychological perspective. *European Journal of Psychology of Education, 14*(1), 23-40. DOI:10.1007/BF03173109

Kremmel, B., & Harding, L. (2020). Towards a comprehensive, empirical model of language assessment literacy across stakeholder groups: developing the Language Assessment Literacy Survey. *Language Assessment Quarterly, 17*(1), 100-120. DOI: 10.1080/15434303.2019.1674855

Kvale, S., & Brinkmann, S. (2009). *InterViews: Learning the craft of qualitative research interviewing* (2nd ed.). SAGE Publications.

Land, S., & Jonassen, D. (Eds.) (2008). *Theoretical foundations of learning environments*. Taylor & Francis Group.

Lane, S., Raymond, M. R., & Haladyna, T. M. (Eds.) (2016). *Handbook of test development.* Routledge.

Lantolf, J. P., & Thorne, S. L. (2006). *Sociocultural theory and the genesis of second language development*. Oxford University Press.

Lave, L. (1988). *Cognition in practice: Mind, mathematics, and culture in everyday life.* Cambridge University Press.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation.* Cambridge University Press.

Leighton, J. P. (2017). *Using think-aloud interviews and cognitive labs in educational research.* Oxford University Press.

Lesgold, A., Rubinson, H., Feltovich, P., Glaser, R., Klopfer, D., & Wang, Y. (1988). Expertise in a complex skill: Diagnosing X-ray pictures. In M. T. Chi, R. Glaser and M. J. Farr (Eds.), *The nature of expertise* (pp. 311–342). Lawrence Erlbaum Associates.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science, 29(3)*, 375–419, DOI:10.1207/s15516709cog 000025

Lumley, T. (1993). Reading comprehension sub-skills: Teachers' perceptions of content in an EAP test. *Melbourne Papers in Language testing, 2*(1), 25-60.

Luoma, S. (2004). *Assessing speaking*. Cambridge University Press.

Lynch, B. K. (2003). *Language assessment and programme evaluation.* Edinburgh University Press.

Mason, J. (2004). Semistructured interview. In M. S. Lewis-Beck, A. Bryman, and T. F. Liao (Eds.), *The SAGE encyclopedia of social science research methods* (pp.1020-1022). SAGE Publications.

Mason, R. (2001). Models of online courses. *Education at a Distance*, *15*(7), 1-14.

Mayes, T. (2001). Learning technology and learning relationships. In J. Stephenson (Ed.), *Teaching and learning online: Pedagogies for new technologies* (pp.16-25). Kogan Page.

Mazuro, C., & Rao, N. (2011). Online discussion forums in Higher Education: Is 'lurking' working? *International Journal for Cross-disciplinary Subjects in Education, 2*(2), 364-371.
DOI: 10.20533/ijcdse.2042.6364.2011.0051

McCrum-Gardner, E. (2008). Which is the correct statistical test to use? *British Journal of Oral and Maxillofacial Surgery, 46*, 38-41. DOI: 10.1016/j.bjoms.2007.09.002

McGraw, K. O., & Wang, S. P. (1996). Forming inferences about some Intraclass correlation coefficients. *Psychological Methods, 1*(1), 30-46.

McLaughlin, B. (1990). Restructuring. *Applied Linguistics*, *11*(2), 113–128.

Meara, P. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjaer, and J. Williams (Eds.), *Performance and competence in second language acquisition* (pp.35-53). Cambridge University press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp.13–104). American Council on Education and Macmillan.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*(3), 241-256.

Morse, J. M. (2012). The implications of interview type and structure in mixed-method designs. In J. F. Gubrium (Ed.), *The SAGE handbook of interview research: The complexity of the craft* (2nd ed.) (pp.193-204). SAGE Publications.

Naeem, N., van der Vleuten, C., & Alfaris, E. A. (2012). Faculty development on item writing substantially improves item quality. *Advances in Health Sciences Education: Theory and Practice*, *17*(3), 369–376. DOI: 10.1007/s10459-011-9315-2

Nesher, P., & Sukenik, M. (1991). The effect of formal representation on the learning of ratio concepts. *Learning and Instruction, 1*(2), 161-175.

Neves, D. M., & Anderson, J. R. (1981). Knowledge Compilation: Mechanisms for the automatization of cognitive skills. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 57–84). Lawrence Erlbaum Associates.

Newby, A. C. (1992). *Training evaluation handbook.* Gower.

Newell, A. (2013). Putting it all together. In D. Klahr and K. Kotovsky (Eds.), *Complex information processing* (pp.399-440). Lawrence Erlbaum Associates, Inc.

Ngo, X. M. (2016, October). Demystifying item writing: the need for a theoretical framework. Paper presented at the *4th British Council New Directions in English Language Assessment*, Hanoi, Vietnam.

Nichols, P., & Fulkerson, D. (2010). *Informing design patterns using research on item writing expertise* (Pearson Large-Scale Assessment Technical Report 9). Retrieved August 25, 2019 from https://www.researchgate.net/publication/268303080_Informing_Design_Patterns_Using_Research_on_Item_Writing_Expertise .

Nitko, A. J. (1984). [Review of the book *A technology for test item writing*]. *Journal of Educational Measurement, 21*(2), 201-204.

Nunan, D. (1988). *The learner-centred curriculum: A study in second language teaching*. Cambridge University Press.

O'Neill, L. D., Mortensen, S. M. R., Nørgård, C., Øvrehus, A. L. H., & Friis, U. G. (2019). Screening for technical flaws in multiple-choice items. A generalizability study. *Dansk Universitetspædagogisk Tidsskrift, 26*, 51-65.

Osterlindt, S. J. (1998). *Constructing test items: Multiple-choice, constructed-response, performance, and other formats.* Kluwer Academic Publishers.

O'Sullivan, B. (2004*). Issues in testing Business English: The revision of the Cambridge Business English certificates.* Cambridge University Press.

O'Sullivan, B., & Weir, C.J. (2011). Test development and validation. In B. O'Sullivan (Ed.), *Language testing: Theories and practices* (pp.13-32). Palgrave Macmillan.

O'Sullivan, B., Dunlea, J., Spiby, R., Westbrook, C., & Dunn, K. (2020). *Aptis general technical manual. Version 2.2* (Report No. TR/2020/001). The British Council. https://www.britishcouncil.org/exam/aptis/research/publications/technical/general-technical-manual-version-2-2Peirce, B. N. (1992). Demystifying the TOEFL Reading Test. *TESOL Quarterly, 26*(4), 665-91.

Phillips, J. J. (1991). *Handbook of training evaluation and measurement methods.* Gulf Publishing Company.

Pill, J., & Harding, L. (2013). Defining the language assessment literacy gap: Evidence from a parliamentary inquiry. *Language Testing, 30*(3), 381-402. DOI: 10.1177/0265532213480337

Popham, W. J. (1994). The instructional consequence of criterion-referenced clarity. *Educational Measurement: Issues and Practice, 13*(4), 15-18. DOI:10.1111/j.1745-3992.1994.tb00565.x

Qin, Y., Sohn, M. H., Anderson, J. R., Stenger, V. A., Fissell, K., Goode, A., & Carter, C. S. (2003). Predicting the practice effects on the blood oxygenation level-dependent (BOLD) function of fMRI in a symbolic manipulation task. *Proceedings of the National Academy of Sciences of the United States of America. 100(*8), 4951–4956. DOI: 10.1073/pnas.0431053100

Proctor, R. W., & Dutta, A. (1995). *Skill acquisition and human performance*. SAGE Publications.

Rapley, T. (2011). Some pragmatics of qualitative data analysis. In D. Silverman (Ed.), *Qualitative research*, 4th ed. (pp. 331-346). SAGE Publications.

Reed, S.K., & Bolstad, C. A. (1991). Use of examples and procedures in problem solving. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 17*(4), 753-766.

Renninger, K., & Hidi, S. (2016). *The power of interest for motivation and engagement*. Routledge.

Richards, J. (2001). *Curriculum Development in Language Teaching*. Cambridge University Press.

Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive Tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review, 14,* 249–255. DOI: 10.3758/BF03194060

Romiszowski, A. (2009). Fostering skill development outcomes. In C.M. Reigeluth and A.A. Carr-Chellman (Eds.), *Instructional-design theories and models: Building a common knowledge base* (pp.199-224). Routledge.

Rossi, O. (2017). Assessment literacy for test writers: What do people who write language tests need to know about testing? Paper presented at the LAEL PG Conference. Lancaster, UK, June 26-27.

Rossi, O., & Brunfaut, T. (2019). Test item writers. In J. I. Liontas (Ed.), *The TESOL Encyclopaedia of English Language Teaching* (pp.1-7). John Wiley & Sons. DOI:10.1002/ 9781118784235.eelt0981

Rossi, O., & Brunfaut, T. (2019). Towards social justice for item writers: Empowering item writers through language assessment literacy training. Pater presented at the 41st LTRC. Atlanta, USA, March 4-8.

Rossi, O., & Brunfaut, T. (2021). Text authenticity in listening assessment: Can item writers be trained to produce authentic-sounding texts? *Language Assessment Quarterly.* DOI: 10.1080/15434303.2021.1895162

Rubin, H. J., & Rubin, I. (2005). *Qualitative interviewing: The art of hearing data* (2nd ed.). SAGE Publications.

Ryan, E., & Brunfaut, T. (2016). When the test developer does not speak the target language: The use of language informants in the test development process. *Language Assessment Quarterly, 13*(4), 393-408. DOI: 10.1080/15434303.2016.1236110

Salisbury, K. (2005). *The Edge of Expertise: Towards an Understanding of Listening Test Item Writing as Professional Practice*. Unpublished doctoral thesis, King's College London, UK.

Salvucci, D. D. (2006). Modeling driver behavior in a cognitive architecture. *Human Factors*, *48*, 362–380. DOI: 10.1518/2F001872006777724417

Saville, N. (2003). The process of test development and revision within UCLES EFL. In C. Weir and M. Milanovic (Eds.), *Continuity and innovation: Revising the Cambridge Proficiency in English examination 1913-2002* (pp.57-120). Cambridge University Press.

Schwier, R., & Daniel, B. (2008). Implications of a Virtual Learning Community model for designing Distributed Communities of Practice in higher education. In C. Kimble, P. Hildreth, and I. Bourdon (Eds.), *Communities of Practice: Creating learning environments for educators* (Vol.2, pp.347-366). Information Age Publishing, Inc.

Scott, K. R., King, A. M., Estes, M. K., Conlon, L. W., Jones, J. S., & Phillips, A. W. (2019). Evaluation of an intervention to improve quality of single-best answer multiple-choice questions. *The Western Journal of Emergency Medicine*, *20*(1), 11–14. DOI: 10.5811/westjem.2018.11.39805

Shin, D. (2012). Item writing and writers. In G. Fulcher and F. Davidson (Eds.), *The Routledge handbook of language testing* (pp.237-248). Routledge.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428.

Singley, M. K., & Anderson, J. R. (1985). The transfer of text-editing skill. *International Journal of Man-Machine Studies, 22*, 403– 423. DOI: 10.1016/S0020-7373(85)80047-X

Smith, S., & Geist, M. (2020). TERM model: The incorporation of mentorship as a test-item improvement strategy. *Teaching and Learning in Nursing, 0,* pp. 1-3. DOI: 10.1016 /j.teln.2020.05.008

Snyder, W. M., & Wenger, E. (2010). Our world as a learning system: A Communities-of-Practice approach. In C. Blackmore (Ed.), *Social learning systems and communities of practice* (pp. 107–124). The Open University. DOI: 10.1007/978-1-84996-133-2_7

Sohn, M. H., Goode, A., Stenger, V. A., Carter, C. S., & Anderson, J. R. (2003). Competition and representation during memory retrieval: Roles of the prefrontal cortex and the posterior parietal cortex. *Proceedings of National Academy of Sciences, 100*(12), 7412–7417. DOI: 10.1073/pnas.0832374100

Spaan, M. (2007). Evolution of a test item. *Language Assessment Quarterly*, 4(3), 279-293, DOI: 10.1080/15434300701462937

Speelman, C., & Kirsner, K. (2005). *Beyond the learning curve: The construction of mind*. Oxford University Press.

Steffe, L.P., & Gale, J. (Eds.) (1995). *Constructivism in education.* Lawrence Erlbaum Associates.

Strauss, S., & Stavy, R. (1982). U-shaped behavioural growth: Implications for theories of development. In W. W. Hartup (Ed.), *Review of Child Development Research* (pp.547-599). The University of Chicago Press.

Sydorenko, T. (2011). Item writer judgements of item difficulty versus actual item difficulty: A case study. *Language Assessment Quarterly, 8*(1), 34-52, DOI: 10.1080/ 15434303.2010.536924

Tashakkori, A., & Creswell, J. W. (2007). Editorial: The New Era of Mixed Methods. *Journal of Mixed Methods Research, 1*(1), 3-7. DOI: 10.1177/2345678906293042

Taylor, K., & Dionne, J. (2000). Accessing problem-solving strategy knowledge: The complementary use of concurrent verbal protocols and retrospective debriefing. *Journal of Educational Psychology, 92*(3), 413-425. DOI: 10.1037/0022-0663.92.3.413

Taylor, L. (2013). Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections. *Language Testing, 30*(3), 403-412. DOI: 10.1177/0265532213480338

Teddlie, C., & Tashakkori, A. (2009). *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences*. SAGE Publications.

Thorndike, E. L., & Woodworth, R. S. (1901). The influence of improvement in one mental function upon the efficiency of other functions. *Psychological Review, 8*, 247– 261.

TOEFL (2018). *TOEFL iBT test framework and test development* (volume 1). Retrieved August 15, 2019 from https://www.ets.org/s/toefl/pdf/toefl_ibt_research_insight.pdf.

Tricio, J., Montt, J., Orsini, C. A., & Ormeño, A. (2018, March). Faculty development as part of a comprehensive quality assurance protocol significantly improves multiple-choice item-writing quality. Poster presented at the *Authentic Assessment across Continuum of Health Professions Education Symposium*, Abu Dhabi, UAE. DOI: 10.13140/RG.2.2.29849.52322

van Moere, A. (2014). Raters and ratings. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp.1358-1374). John Wiley & Sons.

Vaughn, K. W. (1951). Planning the objective test. In E. F. Lindquist (Ed.), *Educational measurement* (pp.159-184). American Council on Education.

Vygotsky, L. S. (1979). *Mind in society: The development of higher psychological process.* Harvard University Press.

Wagner, E. (2016). Authentic texts in the assessment of L2 listening ability. In J. Banerjee and D. Tsagari (Eds.), *Contemporary second language assessment* (pp.103–123). Continuum.

Wagner, E. (2018). A comparison of L2 listening performance on tests with scripted or authenticated spoken texts. In G. J. Ockey and E. Wagner (Eds.), *Assessing L2 listening: Moving towards authenticity* (pp.29–44). John Benjamins.

Weber, K. (2009). The relationship of interest to internal and external motivation. *Communication Research Reports, 20*(4), 376-383. DOI:10.1080/08824090309388837

Weiler, T. (2018). *Investigating the construct tested through four item types used to assess lexicogrammatical competence in English as a foreign language*. Unpublished doctoral thesis, Lancaster University, UK.

Weir, C. J. (1995). *Understanding and developing language tests*. Phoenix.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Palgrave Macmillan.

Weir, C., & Roberts, J. (1994). *Evaluation in ELT.* Blackwell.

Welch, C. (2006). Item and prompt development in performance testing. In S. M. Downing and T. M. Haladyna (Eds.), *Handbook of test development* (pp.303-327). Lawrence Erlbaum Associates.

Welford, A.T. (1968). *Fundamental of skill.* Methuen.

Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity.* Cambridge University Press.

Wenger, E. (2000). Communities of Practice and Social Learning Systems. *Organization*, *7*(2), 225–246.

Wenger, E. (2010). Communities of Practice and Social Learning Systems: The Career of a Concept. In C. Blackmore (Ed.), *Social learning systems and communities of practice* (Vol. 14, pp. 179–198). The Open University. DOI: 10.1007/978-1-84996-133-2_11

Wenger, E. (2012). Communities of practice: Learning, meaning and identity. In J. Jameson and S. D. Freitas (Eds.), *The e-learning reader* (pp. 291–294). Bloomsbury Publishing Plc.

Wenger, E. (2018). A social theory of learning. In K. Illeris (Ed.), *Contemporary theories of learning: Learning theorists in their own words* (2nd ed., pp. 219–228). Taylor & Francis Group.

Wenger, E., McDermott, R. A., & Snyder, W. (2002). *Cultivating Communities of Practice: A guide to managing knowledge.* Harvard Business Review Press.

Wenger-Trayner, E., & Wenger-Trayner, B. (2015). Communities of practice: A brief introduction. Retrieved July 25, 2020 from https://wenger-trayner.com/introduction-to-communities-of-practice/.

Wesman, A.G. (1971). Writing the test item. In R. Thorndike, W. H. Angoff and E. F. Lindquist (Eds.), *Educational measurement* (2d ed.) (pp.99-111). American Council on Education.

Wheelan, C. (2013). *Naked statistics: Stripping the dread from the data*. W. W. Norton.

Yan, X., & Park, H. (2019, March). Many-facet Rasch modeling of rater performance over the course of a rater certification program. Paper presented at the *ILTA Language Testing Research Colloquium*, Atlanta, USA.

Yang, B. (2004). Holistic learning theory and implications for human resource development. *Advances in Developing Human Resources,* *6*(2), 241-262. DOI:10.1177/1523422304263431

Young, M. F., Kulikowich, J. M., & Barab, S. A. (1997). The unit of analysis for situated assessment. *Instructional Science*, *25*(2), 133–150. DOI:10.1023/A:1002971532689

Yurdakul, B., Basokcu, T. O., & Yazicilar, U. (2020). Evaluation of the professional development program for secondary math teachers on item writing related to higher order thinking skills. *Journal of Teacher Education and Educators*, *9*(1), 83–106.

Appendices

Appendix 1: Induction item-writing training course syllabus

Topic	Activities	Materials	Modes of interaction	Types of feedback
Module 1: Introduction to item writing				
Introduction to using the CEFR for item writing	[input] studying a presentation about the CEFR; [practice] applying the CEFR illustrative scale descriptors	PPT, the CEFRTrain project website, worksheet	Individually	Immediate feedback from the CEFRTrain website
	[input/discussion] reading and discussing an article on using the CEFR to produce test items	Davidson & Fulcher (2007), discussion questions	Individually / In groups	Peer-feedback, tutor feedback summary
What makes a good test item?	[input] studying a presentation about the principles of item writing; [comprehension check] doing a follow-up quiz	PPT, online quiz	Individually	Immediate quiz score, tutor feedback summary
Item specifications and quality review (QR) checklists	[input] studying a presentation on the use of specifications and QR checklists in item writing; [practice] creating a QR checklist for a pre-course assignment task	PPT, pre-course assignment specifications, QR checklist template	Individually / In groups	Peer-feedback, tutor feedback summary
Module 2: Writing grammar items				

The construct of grammar assessment	[input] studying a presentation on the construct of grammar assessment and the use of the *Core Inventory* document; [comprehension check] using the *Core Inventory* to complete an online quiz	PPT, the *Core Inventory* document, online quiz	Individually	Immediate quiz score, tutor feedback summary
Introduction to writing multiple-choice (MC) items	[input] studying a presentation on the principles of writing MC items; [practice] identifying flaws in 10 MC grammar items; [practice] reflecting on the MC grammar items produced for the pre-course assignment	PPT, worksheet with 10 MCQ grammar items, grammar items produced for the pre-course assignment	Individually / In groups	Peer-feedback, tutor feedback summary
Grammar item-writing practice	[practice] Producing 3 MC grammar items at 3 different proficiency levels; [practice] giving feedback on items produced by peers against a QR checklist	MCQ grammar item specifications, QR checklist	Individually / In groups	Peer-feedback, individual tutor-feedback, tutor feedback summary
Module 3: Writing vocabulary items				
The construct of vocabulary assessment	[input/discussion] reading and discussing a chapter on the construct of vocabulary assessment	Meara (1996), discussion questions	Individually / In groups	Peer-feedback, tutor feedback summary
Using corpus methods to determine	[input] studying a presentation on vocabulary frequency and on using *Lextutor*;	PPT, K1-K15 vocabulary lists, worksheet with	Individually	Individual tutor-feedback, tutor

vocabulary frequency	[practice] adjusting the lexical complexity of a genuine text to B1 level.	the original text, *Lextutor* website		feedback summary
Introduction to writing multiple-matching items	[input] studying a presentation on the principles of writing multiple-matching tasks; [practice] identifying flaws in 4 multiple-matching vocabulary tasks	PPT, worksheet with 4 multiple-matching vocabulary items	Individually / In groups	Peer-feedback, tutor feedback summary
Vocabulary item-writing practice	[practice] Producing a multiple-matching vocabulary task at B1 level; [practice] giving feedback on tasks produced by peers against a QR checklist	Multiple-matching vocabulary task specifications, QR checklist	Individually / In groups	Peer-feedback, individual tutor-feedback, tutor feedback summary
Module 4: Writing tasks to test productive skills (speaking and writing)				
The construct of speaking assessment	[input] studying a presentation on the construct of speaking assessment and the principles of producing oral interview questions; [practice] producing an OPE schedule [practice] giving feedback on OPEs produced by peers against a QR checklist	PPT, OPE specifications, QR checklist	Individually / In groups	Peer-feedback, individual tutor-feedback, tutor feedback summary
Speaking item-writing practice (picture description and	[input] studying a presentation on the principles of producing picture	PPT, task specifications, QR checklists	Individually / In groups	Peer-feedback, individual tutor-feedback,

short talk speaking tasks)	description and short talk speaking tasks; [practice] producing a picture description task or a short talk task; [practice] giving feedback on tasks produced by peers against a QR checklist			tutor feedback summary
The construct of writing assessment	[input] studying a presentation on the construct of writing assessment and the general principles of producing writing prompts; [practice] analysing and improving on a writing prompt (each trainee in a group is allocated a different prompt)	PPT, worksheet with writing prompts	Individually / In groups	Peer-feedback, tutor feedback summary
Writing item-writing practice	[practice] reflecting on the writing prompt produced for the pre-course assignment; [practice] producing a writing prompt for the 'online social network interaction' writing task; [practice] giving feedback on the writing prompts produced by peers against a QR checklist	Writing prompt specifications, writing prompts produced for the pre-course assignment, QR checklists	Individually / In groups	Peer-feedback, individual tutor-feedback, tutor feedback summary
Module 5: Writing reading tasks				
The construct of reading assessment	[input] studying a presentation on the construct of reading assessment; [practice]	PPT, task worksheet	Individually	The task key provided with the task

	learning to identify reading subskills			
Adapting reading texts to different proficiency levels	[input] studying a presentation on selecting and adapting reading texts, and on checking text readability using *Cohmetrix*; [practice] adapting a reading text of trainee's choice to B2 level	PPT, reading text specifications, *Cohmetrix* website	Individually / In groups	Peer-feedback, tutor feedback summary
Reading item-writing practice	[input] studying a presentation on writing reading tasks (True-False, sentence completion, short-answer questions, rearrangement, information transfer); [practice] writing two tasks of different types for the adapted reading text; [practice] giving feedback on the reading tasks produced by peers against a QR checklist	PPT, reading task specifications, QR checklist	Individually / In groups	Peer-feedback, individual tutor-feedback, tutor feedback summary
Module 6: Writing listening tasks				
The construct of listening assessment; listening texts authenticity	[input] studying a presentation on the construct of listening assessment and listening text authenticity; [practice] reflecting on the authenticity of the listening text produced for the pre-	PPT, listening texts produced for the pre-course assignment	Individually / In groups	Peer-feedback, tutor feedback summary

	course assignment, revising the text to make it more authentic-sounding			
Developing listening texts	[input] studying a presentation on developing listening input texts (exploiting genuine sound files, semi-scripting, scripting); [practice] textmapping a genuine sound file	PPT, a sound file, worksheet 'Gist textmapping procedure'	Individually / In groups	Peer-feedback, tutor feedback summary
Listening item-writing practice	[input] studying a presentation on principles and techniques of producing listening text items; [practice] producing three listening tasks (including texts); [practice] giving feedback on the tasks produced by peers against a QR checklist	PPT, listening task specifications, QR checklist	Individually / In groups	Peer-feedback, individual tutor-feedback, tutor feedback summary

Appendix 2: Information sheets

Participant information sheet: Expert judges

Project Title: The nature of item writing skills and their development: Insights from an induction item writer training course

Name of Researcher: Olena Rossi
Email: o.rossi@lancaster.ac.uk

I am a PhD student at Lancaster University and I would like to invite you to take part in a research study investigating item-writing skills and their development.

Please take time to read the following information carefully before you decide whether or not you wish to take part.

What is the study about?

This study will look into training item writers over a three-month period during an online item-writing training course. The study will investigate item-writing skills and the process of their development.

Why have I been invited?

I have approached you because your qualifications and experience in language assessment make you a good candidate to take the role of an expert judge in the study.

I would be very grateful if you would agree to take part in this study.

What will I be asked to do if I take part?

If you decided to take part, this would involve the following:

- Evaluating items produced by 35 participants of the study. Each participant completed a pre-course and a post-course item-writing assignment consisting of three tasks: a grammar task (two multiple choice items at different proficiency levels), a writing prompt and a listening task. The pre-course and the post-course assignments are exactly the same, but participants were asked to produce a new set of items and not to improve of the pre-course ones. In total, you will review 280 tasks.
- You will not be informed which items were written pre- course and which post-course. Quality review checklists with detailed evaluation criteria will be provided and we will

have an online meeting to discuss the item review process and address any questions you might have.

- The sets of items will be sent to you electronically, and you will complete the evaluation at the time and place suitable for you. You will submit the evaluations electronically by a set deadline (8 weeks).
- Item review should take approximately 10 days of your time and you will be paid £125 per day, £1,250 in total on completion of the item evaluation work.

What are the possible benefits from taking part?

If you take part in this study, you will have a chance to use your assessment expertise to discuss language assessment issues in academic environment. You will also be paid for doing the item evaluation.

Your participation in this study will provide me with insights into the quality of items produced by the participants and will aid me in the investigation of the process of item-writing skill development.

Do I have to take part?

No. It's completely up to you to decide whether or not you take part. Your participation is voluntary.

What if I change my mind?

If you change your mind, you are free to withdraw at any time before or during the data collection stage. If you want to withdraw during the data collection, please let me know, and I will extract any data you contributed to the study and destroy it. Data means the item evaluations you have produced.

Please note that, if you decide to withdraw during the data collection stage, you will not be paid for any item evaluation work you might have done by that time. It will be impossible to withdraw data generated by you once you have submitted you item evaluation judgements and been paid.

What are the possible disadvantages and risks of taking part?

Taking part in this study will entail considerable time investment on your part. I estimate you will spend about 10 days of your time evaluating items produced by the study participants.

Will my data be identifiable?

Only I, the researcher conducting this study, and my supervisor Dr. Tineke Brunfaut will have access to the data generated during the study.

I will keep all personal information about you (e.g. your name and other information about you that can identify you) confidential, that is I will not share it with others. I will anonymise hard copies of any data. This means that I remove any personal information.

How will my data be stored?

Your data will be stored in encrypted files (that is no-one other than me, the researcher, and my supervisor will be able to access them) and on password-protected computers.

I will store hard copies of any data securely in locked cabinets in my office.

I will keep data that can identify you separately from non-personal information (e.g. your views on a specific topic).

In accordance with University guidelines, I will keep the data securely for a minimum of ten years.

How will I use the information you have shared with me and what will happen to the results of the research study?

I will use the data you have shared for academic purposes only. This will include my PhD thesis and other publications, for example journal articles. I may also present the results of this study at academic conferences. The study results may also be used for teaching purposes (e.g. future item-writing courses).

Who has reviewed the project?

This study has been reviewed and approved by the Faculty of Arts and Social Sciences and Lancaster Management School's Research Ethics Committee.

What if I have a question or concern?

If you have any queries or if you are unhappy with anything that happens concerning your participation in the study, please contact myself

Olena Rossi
o.rossi@lancaster.ac.uk
+447857644271
Department of Linguistics and English Language
Lancaster University
Bailrigg, Lancaster LA1 4YW
UK

Or my supervisor

Dr. Tineke Brunfaut
t.brunfaut@lancaster.ac.uk
+441524594084
Department of Linguistics and English Language
Lancaster University
Bailrigg, Lancaster LA1 4YW
UK

If you have any concerns or complaints that you wish to discuss with a person who is not directly involved in the research, you can also contact:

Prof. Elena Semino
Head of Department of Linguistics and English Language
e.semino@lancaster.ac.uk
+441524594176
Lancaster University

Bailrigg, Lancaster LA1 4YW
UK

Thank you for considering your participation in this project.

Participant information sheet: Item-writing trainees

Project Title: The nature of item writing skills and their development: Insights from an induction item writer training course

Name of Researcher: Olena Rossi

Email: o.rossi@lancaster.ac.uk

I am a PhD student at Lancaster University and I would like to invite you to take part in a research study investigating item-writing skills and their development.

Please take time to read the following information carefully before you decide whether or not you wish to take part.

What is the study about?

This study will look into training item writers over a three-month period during an online item-writing training course. The study will investigate item-writing skills and the process of their development.

Why have I been invited?

I have approached you because your position as an assessment consultant as well as your qualifications make you a good candidate to take part in the study and the item writer training course.

I would be very grateful if you would agree to take part in this study.

What will I be asked to do if I take part?

If you decided to take part, this would involve the following:

- Completing an online background questionnaire providing information about your gender, age, nationality, languages spoken, teaching, writing and testing qualifications and experience. The questionnaire will take on average 20 minutes of your time.
- Doing a pre-course item-writing task which consists of writing several test items and should take on average 2 hours of your time. You will have a week to do the task and

will complete it at the time and place suitable for you. After you submit the work electronically it will be evaluated by independent item reviewers.

- If you agree, I will also conduct an online interview with you. However, you do not have to agree to do the interview to take part in the study. During the interview you will be asked questions about the items you wrote. The interview will not take more than 30 minutes. The interview will be audio-recorded.
- You will then take part in a 3-month online training course. During the course I will collect data produced by you, including your responses to quizzes, items produced for practice item-writing tasks, scripts of online group discussions and individual assignments you submit.
- You will be offered 4 feedback questionnaires at various times during the course. You will need to submit the questionnaires online by a set deadline. Each questionnaire will take on average 15 minutes of your time.
- After the course has finished, you will be asked to do a post-course item-writing task which will consist of writing several test items and should take on average 2 hours of your time. You will have a week to do the task and will complete it at the time and place suitable for you. After you submit the work electronically it will be evaluated by independent item reviewers.
- If you agree, I will conduct an online interview with you. However, you do not have to agree to do the interview to take part in the study. During the interview you will be asked questions about the items you wrote. The interview will not take more than 30 minutes. The interview will be audio-recorded.

What are the possible benefits from taking part?

If you take part in this study you will receive thorough professional training and, on successful completion, will receive a certificate of attendance. This will enable you to do work as an item writer. Your participation in this study will also provide me with insights into the process of item-writing skill development.

Do I have to take part?

No. It’s completely up to you to decide whether or not you take part. Your participation is voluntary.

If you decide not to take part in this study, this will not affect your position in the company and your relations with your employer. This will also not affect further professional training opportunities you receive within the organisation.

What if I change my mind?

If you change your mind, you are free to withdraw at any time before the online course begins. If you want to withdraw, please let me know, and I will extract any data you contributed to the study and destroy it. Data means the information, views, ideas, etc. that you and other participants will have shared with me.

If you decide to withdraw from the study during the course, you will have to quit the course as this cohort is run for the purpose of this study. In this case I will extract any data you contributed to the course and destroy it. Data means your participation in group discussions (in some cases this will mean deleting the entire group discussion), item-writing tasks you have done, quizzes you have submitted and any other way data generated from your participation in the course.

If you withdraw from this cohort, you will still have a chance to do the course later and will be put on a waiting list.

Please note that it will be impossible to withdraw data generated by you once the data analysis has started 2 weeks after the end of the data collection.

What are the possible disadvantages and risks of taking part?

Taking part in this study will entail substantial time investment on your part. At the initial stage, you will need to spend 2 – 2.5 hours of your time completing a questionnaire, doing a pre-course task and, possibly, an interview. The course will run for 3 months and will require you to spend 2-4 hours per week doing tasks and activities. After you complete the course you will spend 2 – 2.5 hours doing a post-course tasks and, possibly, an interview.

Will my data be identifiable?

Only I, the researcher conducting this study, and my supervisor Dr. Tineke Brunfaut will have access to the data generated during the study.

I will keep all personal information about you (e.g. your name and other information about you that can identify you) confidential, that is I will not share it with others. I will anonymise transcripts of audio recordings and hard copies of any data. This means that I remove any personal information.

How will my data be stored?

Your data will be stored in encrypted files (that is no-one other than me, the researcher, and my supervisor will be able to access them) and on my password-protected computer.

I will store hard copies of any data securely in locked cabinets in my office.

I will keep data that can identify you separately from non-personal information (e.g. your views on a specific topic).

In accordance with University guidelines, I will keep the data securely for a minimum of ten years.

How will I use the information you have shared with me and what will happen to the results of the research study?

I will use the data you have shared for academic purposes only. This will include my PhD thesis and other publications, for example journal articles. I may also present the results of this study at academic conferences. The study results may also be used for teaching purposes (e.g. future item-writing courses). I will inform policy-makers within your organisation about the results of this study, but only as a whole (without sharing your identities or any personal information about you).

When writing up the findings from this study, I would like to reproduce some of the views and ideas you shared with me. When doing so, I will only use anonymised quotes (e.g. from my interview with you), so that although I will use your exact words, you cannot be identified in our publications.

Who has reviewed the project?

This study has been reviewed and approved by the Faculty of Arts and Social Sciences and Lancaster Management School's Research Ethics Committee.

What if I have a question or concern?

If you have any queries or if you are unhappy with anything that happens concerning your participation in the study, please contact myself

Olena Rossi
o.rossi@lancaster.ac.uk
+447857644271
Department of Linguistics and English Language
Lancaster University
Bailrigg, Lancaster LA1 4YW
UK

Or my supervisor

Dr. Tineke Brunfaut
t.brunfaut@lancaster.ac.uk
+441524594084
Department of Linguistics and English Language
Lancaster University
Bailrigg, Lancaster LA1 4YW
UK

If you have any concerns or complaints that you wish to discuss with a person who is not directly involved in the research, you can also contact:

Prof. Elena Semino
Head of Department of Linguistics and English Language
e.semino@lancaster.ac.uk
+441524594176
Lancaster University
Bailrigg, Lancaster LA1 4YW
UK

Thank you for considering your participation in this project.

Appendix 3: Consent forms

Consent form: Expert judges

Project Title: The nature of item writing skills and their development: Insights from an induction item writer training course

Name of Researcher: Olena Rossi

Email: o.rossi@lancaster.ac.uk

Please tick each box from 1 to 6

1. I confirm that I have read and understand the information sheet for the above study. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily. □
2. I understand that my participation is voluntary and that I am free to withdraw at any time during my participation in this study and within 2 weeks after I took part in the study, without giving any reason. If I withdraw within 2 weeks of taking part in the study my data will be removed. □
3. I understand that any information given by me may be used in future reports, academic articles, publications or presentations by the researcher/s, but my personal information will not be included and I will not be identifiable. □
4. I understand that my name will not appear in any reports, articles or presentations without my consent. □
5. I understand that data will be kept according to University guidelines for a minimum of 10 years after the end of the study. □
6. I agree to take part in the above study. □

___________________________	_________________	__________________
Name of Participant	Date	Signature

I confirm that the participant was given an opportunity to ask questions about the study, and all the questions asked by the participant have been answered correctly and to the best of my ability. I confirm that the individual has not been coerced into giving consent, and the consent has been given freely and voluntarily.

Signature of Researcher /person taking the consent__________________________

Date ___________ Day/month/year

One copy of this form will be given to the participant and the original kept in the files of the researcher at Lancaster University

Lancaster University

Consent form: Item-writing trainees

Project Title: The nature of item writing skills and their development: Insights from an induction item writer training course

Name of Researcher: Olena Rossi

Email: o.rossi@lancaster.ac.uk

Please tick each box from 1 to 7

1. I confirm that I have read and understand the information sheet for the above study. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily. □
2. I understand that my participation is voluntary and that I am free to withdraw at any time during my participation in this study and within 2 weeks after I took part in the study, without giving any reason. If I withdraw within 2 weeks of taking part in the study my data will be removed. □
3. I understand that any information given by me may be used in future reports, academic articles, publications or presentations by the researcher/s, but my personal information will not be included and I will not be identifiable. □
4. I understand that my name will not appear in any reports, articles or presentations without my consent. □
5. I understand that any interviews will be audio-recorded and transcribed and that data will be protected on encrypted devices and kept secure. □
6. I understand that data will be kept according to University guidelines for a minimum of 10 years after the end of the study. □
7. I agree to take part in the above study. □
8. I agree to take part in pre-course and post-course online interviews. I understand that I can refuse to do the interviews and can still participate in the study. □

_____________________________ _________________ __________________

Name of Participant Date Signature

I confirm that the participant was given an opportunity to ask questions about the study, and all the questions asked by the participant have been answered correctly and to the best of my ability. I confirm that the individual has not been coerced into giving consent, and the consent has been given freely and voluntarily.

Signature of Researcher /person taking the consent__________________________

Date ___________   Day/month/year

One copy of this form will be given to the participant and the original kept in the files of the researcher at Lancaster University

Appendix 4: Item-writing assignment instructions

This activity is being carried out as part of a study into the nature of item-writing skills and their development.

You can do the tasks at a time and place suitable for you. While doing the tasks, please take a note of how much time it took you to write each item. This does not suggest you need to do the task as quickly as possible – please feel free to use as much or as little time as you personally need within the timeframe provided. Please also note you do not have to complete the task in one go and can return to it several times during the week. In this case, please remember to add up the time it took you to write the items.

To complete some of the tasks you will need to use the *Core Inventory for General English* document which specifies grammar exponents, functions and topics at different CEFR levels. You will find the document attached to this email. You can find more information about the CEFR at http://www.coe.int/t/dg4/linguistic/cadre1_en.asp You will also need to comply with lexical level specifications by using an online tool http://www.lextutor.ca/vp/ Please study a short tutorial attached to this email on how to access the BNC (British National Corpus) on *Lextutor.*

Please complete the three tasks within 7 days and return this document via email to o.rossi@lancaster.ac.uk. Please name the document *pre-course task_name_surname* (e.g. *pre-course task_olena_rossi*).

Thank you!

Task 1: Grammar items

for a general English proficiency test, adult candidates, unspecified nationality

Please write two multiple-choice items: one item at CEFR[12] A2 level and one item at CEFR C1 level. While writing the items, please follow the specifications below:

Specifications	A2	C1
Task description	Sentence completion based on the appropriacy of grammatical meaning and/or form	Sentence completion based on the appropriacy of grammatical meaning and/or form
Format	3-option multiple choice	3-option multiple choice
# items	1	1
Word count – stem (including the key)	Max. 10 words	Max. 15 words
Word count - options	1-3 words	1-3 words
Key	Indicate with *	Indicate with *

[12] Common European Framework for Reference, http://www.coe.int/t/dg4/linguistic/cadre1_en.asp

Lexical level	K1[13]	K1 to K5[14]
Grammar (exponent)[15]	Wh-questions in the past	Wish/if only & regrets
Topic	Appropriate at A2	Appropriate at C1
Function	Appropriate at A2	Appropriate at C1

Example of a CEFR A1 level item

A1 item : 59-60 Questions	
Topic:	194 Shopping
Function:	4 Understanding and using prices
Stem	How much _______ ?
Option 1	is the apple*
Option 2	the apple is
Option 3	the apple costs

Please use the templates below to write your items. Make sure to fill in the information about the topic and function (see the *Core Inventory* for the lists of topics and functions appropriate at each CEFR level).

A2 item : 61 Wh-questions in the past	
Topic:	
Function:	
Stem	
Option 1	
Option 2	
Option 3	
Please indicate here how much time it took you to write this item:	

C1 item : 93 Wish/if only & regrets

[13] 1,000 most frequently used words in British English, http://www.lextutor.ca/vp/

[14] 5,000 most frequently used words in British English, http://www.lextutor.ca/vp/

[15] See lists of level-appropriate topics, functions and grammar exponents in the *Core Inventory for General English* document attached to this task.

Topic:	
Function:	
Stem	
Option 1	
Option 2	
Option 3	
Please indicate here how much time it took you to write this item:	

Task 2: Writing prompt

for a general English proficiency test, adult candidates, unspecified nationality

Please write a prompt for an e-mail writing task. While creating the prompt, please follow the specifications below:

Skill focus	A writing task requiring a paragraph-level writing in the form of a formal e-mail, in response to a prompt (an e-mail or a notice).
Task level	B2
Task description	The candidate writes a formal e-mail in response to the task prompt which contains a short e-mail or a notice. The response is a formal e-mail to an unknown reader connected to the information in the prompt (management, customer services, etc).
Instructions to candidates	An **e-mail message/ notice** is presented as the starting point for the e-mail response to be produced. The e-mail message/notice will present a problem / issue / offer / opportunity which the candidate is expected to discuss. An **instruction** is given for the e-mail response. The instruction will specify the intended reader and the purpose/function of the e-mail (complaining, suggesting alternatives, giving advice). All instructions should include the following information: "Write 120-150 words. You have 20 minutes." **See an example of instructions and input e-mail below.**
Length of input e-mail/notice	40-60 words
Overall length of the prompt	80 – 120 words (including the input e-mail/notice)

Grammar of input e-mail/notice and instruction	A1 to B1
Lexis of input e-mail/notice and instruction	K1 to K4[16]

Example

You work for a computer company. This morning you received the following e-mail:

Dear colleagues,

We would like to inform you that from next week, the coffee breaks will be reduced to two a day. Also, because of the high cost of the current machine, we will be replacing it with one that only has regular coffee.

Please feel free to contact us for any feedback.

Kind regards,

The Management Team

Write an e-mail to the Management Team. Fist explain your disagreement with the decision. Then suggest possible alternatives. Write 120-150 words. You have 20 minutes.

Please use the template below to write your prompt

Instructions to candidates:

Input email:

Instructions to candidates (continued):

Please indicate here how much time it took you to write this item:

[16] 4,000 most frequently used words in British English, http://www.lextutor.ca/vp/

Task 3: Listening task

for a general English proficiency test, adult candidates, unspecified nationality

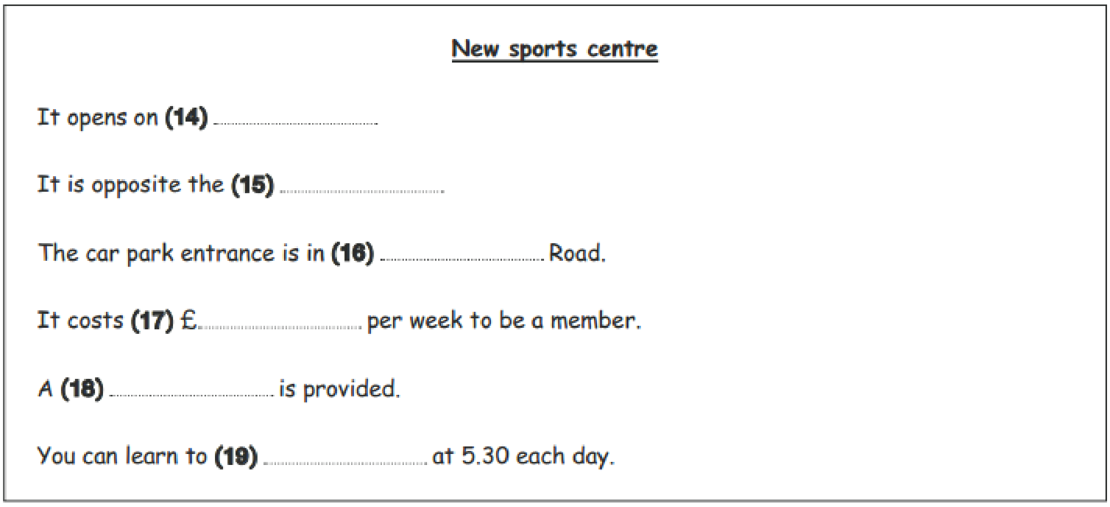
Please write a listening comprehension task at B1 proficiency level. While writing the task, please follow the specifications below:

Task description	Gap-fill
Skill focus	ability to locate and record specific information from a text
Task level	B1
More information about the task	Candidates have a set of notes or sentences, summarising the key content of the text, from which six pieces of information have been removed. As they listen, they fill in the numbered gaps with words from the text which complete the missing information. This may be key pieces of information about places and events, or people talking about courses, trips, holiday activities or other types of factual information. The words candidates need to complete the gaps are heard on the recording: single words, numbers or very short noun phrases.
Instructions to candidates	You will hear … (specify the speakers and the situation., e.g. *a woman talking on the radio about a new sports centre*). For each question, fill in the missing information in the numbered space with a maximum of 3 words or a number.
Listening input specifications	
Text type	A monologue
Text length	max. 300 words
Lexical level	K1 to K3 [17]
Grammatical level	A1 to B1
Topic	From the list of topics for B1 level
Text genre	A monologue: recorded instructions, lectures/presentations, public announcements, TV/radio programmes, short talks, news reports, etc.
Text authenticity	The text should sound like authentic spoken English and not a written script read out. To achieve the authenticity item writers are recommended to write a monologue plan, record an audio version of the text and then transcribe it.
Function	From the list of functions for B1 level

[17] 3,000 most frequently used words in British English, http://www.lextutor.ca/vp/

Item specifications	
Item type	Gap-fill, each gap to be filled with a maximum of 3 words or a number heard in the text. The items are either a set of notes or sentences. Items should follow the order of the text.
Distractors	Distractors will be used in the input text. Each item (except for proper names that are spelt out) should have 1 or 2 distractors.
Items per task	6 in total
Stem length	Maximum 10 words including the key; the stem should not literally repeat what is heard in the text but should be a paraphrase
Stem lexical level	K1 to K2[18]
Stem grammatical level	A1-A2
Response type	Concrete information
Response length	Maximum 3 words or a number from the text
Response lexical level	K1 – K2 (except for proper names that are spelt out, there should be no more than 1 item of this kind per task).

Listening task example:



New sports centre

It opens on **(14)** ....

It is opposite the **(15)** ....

The car park entrance is in **(16)** .... Road.

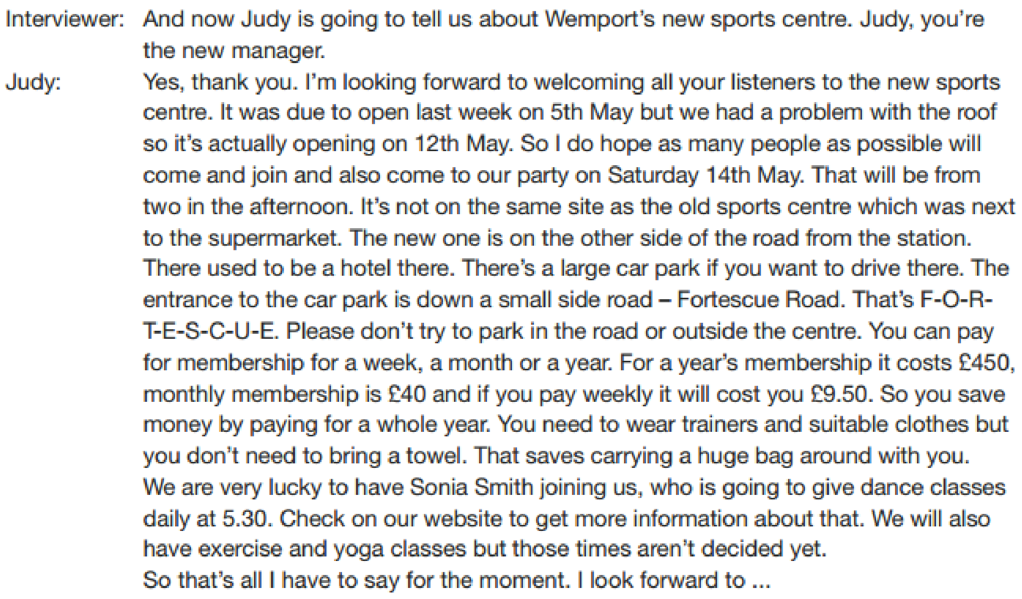
It costs **(17)** £.............................. per week to be a member.

A **(18)** .... is provided.

You can learn to **(19)** .... at 5.30 each day.

Listening text example:

[18] 2,000 most frequently used words in British English, http://www.lextutor.ca/vp/



Interviewer: And now Judy is going to tell us about Wemport's new sports centre. Judy, you're the new manager.

Judy: Yes, thank you. I'm looking forward to welcoming all your listeners to the new sports centre. It was due to open last week on 5th May but we had a problem with the roof so it's actually opening on 12th May. So I do hope as many people as possible will come and join and also come to our party on Saturday 14th May. That will be from two in the afternoon. It's not on the same site as the old sports centre which was next to the supermarket. The new one is on the other side of the road from the station. There used to be a hotel there. There's a large car park if you want to drive there. The entrance to the car park is down a small side road – Fortescue Road. That's F-O-R-T-E-S-C-U-E. Please don't try to park in the road or outside the centre. You can pay for membership for a week, a month or a year. For a year's membership it costs £450, monthly membership is £40 and if you pay weekly it will cost you £9.50. So you save money by paying for a whole year. You need to wear trainers and suitable clothes but you don't need to bring a towel. That saves carrying a huge bag around with you.
We are very lucky to have Sonia Smith joining us, who is going to give dance classes daily at 5.30. Check on our website to get more information about that. We will also have exercise and yoga classes but those times aren't decided yet.
So that's all I have to say for the moment. I look forward to ...

Please use the template below to write your listening task:

Topic	
Function	
Instructions to candidates	
Text	
Items	
Stem 1	
Stem 2	
Stem 3	
Stem 4	
Stem 5	
Stem 6	
Key *(please provide all versions, if there is more than one possible answer)*	
Gap 1	
Gap 2	

Gap 3	
Gap 4	
Gap 5	
Gap 6	

Please indicate here how much time it took you to write this item:

Appendix 5: Item-writing trainee background questionnaire

BACKGROUND QUESTIONNAIRE

This is a short questionnaire about your background, qualifications and experience. It will help us to obtain some relevant information to build a course participant profile and better understand your training needs. It will also serve as a sort of 'getting to know each other' activity – we will collate the information and post the highlights in a module summary on Edmodo (no individual names will be mentioned in the summary - all information will be collated and anonymized).

Your biodata

1 What is your first name?

____

2 What is your surname?

____

3 What is your gender?

○ Male

○ Female

4 How old are you? (in years)

____

Languages you know / use

5 What is your first language(s)?

____

6 If English is your first language, please indicate which country variety it is

○ England

○ Scotland

○ Wales

○ Ireland / Northern Ireland

○ USA

○ Canada

○ Australia

○ New Zealand

○ South African Republic

○ other (please specify) ____

7 What language(s) other than your first language(s) can you use? Please write the name(s) of the language(s) below and indicate your level of proficiency for each skill

Your foreign languages	Speaking ability			Writing ability			Reading comprehension			Listening comprehension		
	E	I	A[19]	E	I	A	E	I	A	E	I	A
#1												
#2												
…												

[19] Elementary, Intermediate, Advanced

Your educational background

8 What degree(s) have you obtained? Please tick all that apply

□ Bachelor's degree (please write the full name of your degree)

____

□ Master's degree (please write the full name of your degree)

____

□ Other (please specify)

____

□ None

9 What ESL/EFL teaching qualification(s) do you have? Please tick all that apply

□ CELTA / Cert. TESOL

□ DELTA / Dip. TESOL

□ PGCE

□ Other (please specify)

____

□ None

10 Do you have any other educational qualifications? Please specify

____

Your ESL / EFL teaching experience

11 How many years ESL/EFL teaching experience do you have?

____

12 What experience do you have teaching ESL/EFL abroad (in countries other than the country you grew up in)?

○ This is my first ESL/EFL job abroad (in the text box, please specify the country and how many years you've worked here, e.g. China, 3 years)

____

○ I've had ESL/EFL jobs in 2 countries (in the text box, please specify the countries and how many years you worked in each)

____

○ I've had ESL/EFL jobs in 3 or more countries (in the text box, please specify all the countries and how many years you worked in each)

____

○ None

13 What ESL/EFL courses have you taught (in any country, including the country you grew up in)? Please choose all options that are true for you.

□ General English to adults (17 years old or more)

□ General English to young learners and teenagers (aged from 1 to 16)

□ EAP (English for Academic Purposes)

□ Business English

□ ESP (English for Specific Purposes - other than Business English)

□ Exam preparation , e.g. for IELTS, FCE, TOEFL, etc.

____

□ Other (please specify)

□ None

Your writing experience

14 Have you ever written any of the following PUBLISHED materials? Please tick all that apply. In the text box next to each option please tell us more about the published materials you have written, e.g. what materials they were, when and where they were published, etc.

□ ESL/EFL textbooks

____

□ Other ESL/EFL teaching materials

____

□ Non-ESL/EFL educational materials

____

□ Article(s) in a magazine / journal

____

□ Fiction / poetry ____

□ Other (please specify)

____

□ No

15 Have you ever written any materials that were read / used by others but NOT PUBLISHED? E.g., teaching materials for your school, blog entries, etc.

○ Yes (please specify what materials they were, who they were written for, when/where/how they were used)

____

○ No

Your ESL / EFL testing experience

16 Do you have experience of classroom assessment?

○ Yes (please indicate the number of years)

____

○ No

What classroom assessment experience do you have? Please tick all that apply

□ testing speaking

□ testing writing

□ testing receptive skills (listening and reading)

□ testing grammar / vocabulary

□ Other (please specify)

____

17 Do you have experience working as an ESL/EFL examiner for an organisation that administers large-scale language exams?

○ Yes (please indicate the number of years)

____

○ No

What examining experience do you have? Please tick all that apply

□ speaking examiner

□ writing examiner

□ other (please specify)

____

18 Do you have experience examining languages other than English?

○ Yes (please specify the languages)

____

○ No

Your ESL / EFL test writing experience

19 Do you have experience writing ESL/EFL tests?

○ Yes (please indicate the number of years)

____

○ No

Pleases specify what kinds of test they were. Please tick all that apply

□ for a professional exam board (please specify the tests and exam board) ____

□ for my school / university / college / language centre (please specify what kinds of test they were, what organisation they were written for, who they were used by, etc.) ____

□ other (please specify) ____

20 Do you have experience writing tests for other languages?

○ Yes (please indicate the languages and the experience you have in years) ____

○ No

Please specify what kinds of test they were. Please tick all that apply

□ for a professional exam board (please specify the tests and exam board) ____

□ for my school / university / college / language centre (please specify what kinds of test they were, what organisation they were written for, who they were used by, etc.) ____

□ other (please specify) ____

21 Have you ever received training in writing language tests? (NOT including the current course)

○ Yes (please tell us more: Who delivered the training? How long was it? Was it held face-to-face or online? What kinds of activity did you do?)
____

○ No

Appendix 6: Item evaluation scales

Item evaluation scales: A2 and C1 grammar items

Objectively-scored criteria		
	Evaluation criteria	Rating scales
G1	Stem: max. 10 (A2) / 15 (C1) words including the key	2 –stem is max. 10 (A2) / 15 (C1) words including the key; 1 – stem is 1-3 words over the limit AND is easy to reduce without much change to the stem; 0 – stem is more than 3 words over the limit OR up to 3 words over the limit but is not possible to reduce without much change to the stem.
G2	Stem: contains one gap only	2 – the stem contains one clear gap to be filled with the correct option; 1 – the stem contains a gap but this is unclear / ambiguously presented; 0 – the stem has more than one gap / does not contain a gap, possibly because the format has been misunderstood.
G3	Options: 3 including the key and distractors	2 – 3 options including the key; 1 – 4 or more options including the key; 0 – less than 3 options OR none of the options can serve as the key.
G4	Options: 1-3 words	2 – each option max. 3 words; 1 – one of the options is over 3 words but is easy to reduce; 0 – two or more options are over 3-word limit OR the over-length option(s) are not possible to reduce without much change to the option(s).
G5	Options: there are no words at the beginning or the end of all options which can be integrated into the stem	2 – no repeating words which can be integrated into the stem in the options; 1 – up to 1 word repeated in the options and can be integrated into the stem without making it over-length; 0 – 2 or more words are repeated in the options AND/OR the repeating word cannot be integrated into the stem without making it over-length.
G6	Key: indicated with asterisk	2 – key is indicated with * 0 – key is not indicated with *
G7	Lexis: K1 (A2) / K1-5 (C1)	2 – all lexis is appropriate, i.e. K1 (A2) / K1-5 (C1);

		1 – one word is above the stated norm; 0 – more than one word is above the stated norm.
G8	Topic: appropriate at A2 / C1 level	2 – topic stated is appropriate for the level and is accurately reflected in the item; 1 – topic stated is appropriate BUT is not accurately reflected in the item; 0 – topic stated is not in the list for the level.
G9	Function: appropriate at A2 / C1 level	2 – function stated is appropriate for the level and is accurately reflected in the item; 1 – function stated is appropriate for the level BUT is not accurately reflected in the item; 0 – function stated is not in the list for the level.
G10	Spelling / grammar / punctuation of the stem and options: correct	2 – all spelling, grammar and punctuation is correct in the stem and options; 1 – one grammar, spelling or punctuation error; 0 – more than one grammar, spelling or punctuation error.

Subjectively-scored criteria

	Evaluation criteria	**Rating scales**	**Guidance for reviewers**
G11	Stem: provides enough context to ensure that the intended construct is tested, including restricting the number of possible correct answers	2 – the stem provides enough context to test the intended construct AND there is only one correct answer in the options; 1 – the stem might be somewhat unclear BUT still provides enough context to test the construct; 0 – the stem does not provide enough context to test the intended construct AND/OR multiple correct answers are possible.	Does the stem provide enough context to test the construct, e.g., if the construct is using Past Tense, is it clear from the stem? Are multiple correct answers possible?
G12	Distractors: strong, plausible	2 – each distractor is strong 1 – one of the distractors is weak 0 – both distractors are weak	Are the distractors plausible? Will the students who have mastered the grammar point tested have more chance to answer correctly? Will the distractors work well in differentiating between weak and strong students? Is it possible to discard any of the distractors without having mastered the grammar point tested?

G13	Distractors: not grammatically correct within the stem	2 – both distractors are not correct within the stem in any of the major English varieties; 1 – one of the distractors might be correct in the slang/dialect English sense, but this might affect only a small part of the candidate population; 0 – one of the distractors is correct in any of the major English varieties OR both distractors might be correct in the slang/dialect English sense, but this might affect only a minor part of the candidate population.	Can the distractors be eliminated as incorrect within the stem by a student who has mastered the grammar point tested? Might one or both distractors be correct in any of the major English varieties? Might one or both distractors be correct in a slang/dialect English variety?
G14	Distractors: grammatically correct as a stand-alone	2 – each distractor is grammatically correct as a stand-alone; 1 – one of the distractors is ungrammatical as a stand-alone; 0 – both distractors are ungrammatical as a stand-alone.	
G15	Key: does not stand out from the distractors	2 – key and both distractors look similar, the key doesn't stand out as different; 1 – key is considerably different from one of the distractors; 0 – key is considerably different from both distractors.	Does the key look considerably different from the distractors? Is the key considerably different from one of the distractors or both?
G16	Grammar exponent: directly targeted in the item	2 – item directly targets the exponent and covers all/most important of its aspects; 1 – item loosely targets the grammar exponent or only a minor aspect of the exponent is targeted; 0 – item does not target any aspect(s) of the grammar exponent.	Does the item directly target the grammar point tested? Does it target the complete or the most important aspect of the grammar point tested?
G17	Grammar of the stem / key: 'standard' English,	2 – the stem and the key are written with 'standard' English usage in mind;	Is the stem sentence, together with the key, written in 'standard' English?

	i.e. not dialect, jargon, etc.	1 – the stem and/or the key might have minor deviations from 'standard' English that will not affect the tested construct; 0 – the stem and/or the key are written in a variety of non-standard English or contain jargon, slang, regional colloquial usage.	Are there any colloquialisms, jargon, slang, non-standard or regional usage?
G18	Content: appropriate, culturally unbiased, not disturbing, suitable for a general-purpose test (i.e. not a specific purpose test)	2 – item is appropriate for international adult candidates, does not have any culturally biased content, will not emotionally disturb candidates, is suitable for a general-purpose test; 1 – content might be somewhat inappropriate for a minority of the candidates because of slight cultural bias, being potentially slightly disturbing AND/OR being more suitable for a specific purposes test; 0 – topic is inappropriate because it deals with religion, violence, abuse, negative emotions (death, divorce, or other disturbing topics), directly discusses controversial political issues AND/OR is culturally biased and will not be clear to candidates who are unfamiliar with the culture of English-speaking countries AND/OR is unsuitable for a general-purpose test.	The items have been written for an adult international audience: Is the content of the items appropriate for international candidates? Might the content of the items affect the performance of (some of) the candidates because of cultural bias? Is the content of the item inappropriate because it is offensive or (culturally) insensitive? E.g. it deals with religion, violence, abuse, controversial political issues, might provoke a memory of negative events in a candidate's life. Is the content of the item suitable for a general-purpose test as opposed to a specific purpose test such as EAP, ESP, etc.?
OA	Overall acceptability of the item for inclusion in a test	2 – on the whole, the item can be accepted in its present form OR after minor revision; 1 – on the whole, the item requires major revision to be accepted; 0 – on the whole, the item should be rejected.	Can the item be accepted as it is? Does the item require any revision? Is it just minor revision or major one? Is it at all possible to revise the item for it to be accepted, or should it be rejected, and a completely new item written?

Item evaluation scales: B2 writing prompts

Objectively-scored criteria

	Evaluation criteria	Rating scales
W1	Input message: 40-60 words	2 – input message is 60 words max; 1 – input message is up to 10 words over or under the word limit AND is easy to reduce/expand without much change to the message; 0 – input message is more than 10 words over or under the limit OR up to 10 words over the limit but cannot be reduced without much change to the message.
W2	Overall length of the prompt: 80-120 words	2 – prompt overall length is 120 words max; 1 – overall, the prompt is up to 10 words over the limit AND is easily reduced without much change to the prompt; 0 – overall, the prompt is more than 10 words over the limit OR up to 10 words over the limit but cannot be reduced without much change to the prompt.
W3	Grammar: A1 – B1	2 – all grammar is appropriate, i.e. A1-A2; 1 – one or two grammar structures are above the stated norm; 0 – more than two grammar structures are above the stated norm.
W4	Lexis: K1 - K4	2 – all lexis is appropriate, i.e. K1-K4; 1 – up to 2 words are above the stated norm; 0 – more than two words are above the stated norm.
W5	Spelling / grammar / punctuation: correct	2 – all spelling, grammar and punctuation of the prompt is correct; 1 – up to two errors; 0 – more than two errors.

Subjectively-scored criteria

	Evaluation criteria	Rating scales	Guidance for reviewers
W6	Input message: a formal email / public notice	2 – input message is a FORMAL email or a PUBLIC notice written according to the rules of the genre (format, style);	Is the input message an email or a notice? If it is an email, is it a FORMAL email?

		1 – input message is a formal email / public notice BUT there are minor violations of the genre format and/or style; 0 – input message is not a formal email / public notice OR is an informal email OR is an attempt at a formal email / public notice but with obvious violations of the genre format and/or style	If it is a notice, is it a PUBLIC notice? Was the email/notice written according to the rules of the genre (format and style)? Are the genre violations slight or major, if any?
W7	Input message: clear and unambiguous	2 – input message is clear, unambiguous, and will facilitate candidates to respond appropriately; 1 – input message is mostly clear BUT there might be some ambiguity in (a) minor detail(s) which will have no effect on candidates' response; 0 – input message is not sufficiently clear and might lead to misinterpretations by candidates affecting their response.	Is the input message clear to candidates in their ability to understand and respond appropriately? Can candidates misinterpret the input message in any way? Is there ambiguity about a minor detail or can it lead to major misinterpretations? Might the misinterpretations affect candidates' response?
W8	Input message: suitable for testing, i.e. NOT a parody, not silly, humorous, sarcastic, etc.	2 – input message is suitable for testing purposes, i.e. not a parody, not silly humorous, sarcastic, or anything else that would be considered unsuitable in a testing situation; 1 - input message is humorous BUT this will not have any negative effect on the testing outcomes; 0 - input message is unsuitable for testing purposes, e.g. is a parody, is silly, humorous, sarcastic, or anything else that would be considered unsuitable in a testing situation.	Is the input message suitable for a testing situation? Does it contain any humour, sarcasm or any other connotations that might be unsuitable in a test? Will the testing outcomes be affected because of the inappropriate input message?
W9	Input message: presents a plausible problem / issue / offer / opportunity which the candidate is expected to discuss	2 – central topic of the input message is a plausible problem, issue, offer or an opportunity; 1 – input message contains a problem, issue, offer or an opportunity BUT this is not central to the message AND/OR is not plausible, i.e. not likely to be encountered in a real-life situation;	Does the input message contain a problem, issue, offer or an opportunity? Is the problem, issue, offer or an opportunity plausible? Is it similar to problems / offers / opportunities candidates will encounter in real-life situations?

		0 – input message does not contain a problem, issue, offer or an opportunity.	Is the problem, issue, offer or an opportunity central to the message?
W10	Instruction: specifies the intended reader of the response email	2 – instruction clearly specifies the intended reader of the response email; 1 – instruction mentions the intended reader BUT this is not sufficiently clear; 0 – instruction does not specify the intended reader of the response email.	
W11	Instruction: specifies the purpose of the response email: complaining, suggesting alternatives, offering advice.	2 – instruction clearly specifies the purpose of the response email: complaining, suggesting alternatives, and/or offering advice; 1 – instruction mentions the purpose of the response email, and the purpose is complaining, suggesting alternatives, and/or offering advice BUT this might not be sufficiently clear AND/OR the purpose specified might not logically follow from the input message; 0 – there is no mention of the purpose of the response email in the instruction OR the purpose of the response email is not complaining, suggesting alternatives, or offering advice.	
W12	Instruction: the purpose of the response email is plausible, i.e. the test-taker is asked to write a response for a plausible reason	2 – the purpose of the response email is plausible, i.e. candidates might expect to write for such a purpose in real-life situations; 1 – the purpose of the response email is not fully plausible; 0 – the purpose of the response email is implausible, i.e. candidates will not write for such a purpose in real-life situations.	Is the purpose of the response email fully plausible? Will the candidate write for such a purpose in real-life situations?
W13	Instruction: the purpose of the response email is	2 - the purpose of the response email is suitably specific, does not allow too much freedom to	Is the purpose of the response email specific enough? Are candidates allowed too much freedom in their responses?

	not too general and does not allow so much freedom to candidates as to result in vastly different responses	candidates and will not create so much variation across the candidate population resulting in possible unreliable ratings; 1 - the purpose of the response email is not specific enough and might create some variation across the candidate population; 0 - the purpose of the response email is too general, will allow too much freedom to candidates and will create a lot of variation across the candidate population resulting in possible unreliable ratings.	Will variations in candidate responses be so much as to result in unreliable ratings?
W14	Instruction: clear and unambiguous, not too wordy or excessive; includes the following information: "Write 120-150 words. You have 20 minutes."	2 – instruction is sufficient, clear, NOT too wordy or excessive, and will facilitate candidates to respond appropriately; 1 – instruction is mostly clear but there might be some ambiguity in (a) minor detail(s) OR a minor detail missing OR the instruction is unnecessarily wordy / excessive; 0 – instruction is not sufficiently clear and might lead to misinterpretations by candidates affecting their response AND/OR an important detail is missing (e.g. "Write 120-150 words / You have 20 minutes.")	Is the instruction clear? Is there any ambiguity? Does it concern (a)minor detail(s) or the whole instruction? Does the instruction contain all necessary information to facilitate candidate response? Are any important details missing from the instruction? E.g. "Write 120-150 words / You have 20 minutes."
W15	Intended response: the task encourages an original response and NOT copying from the input message	2 – the task encourages an original response AND does not allow copying from the prompt; 1 – the task might encourage some copying from the prompt OR some reformulation of the input message; 0 – the task encourages copying OR reformulation from the prompt.	Is the response purpose original (i.e. does not overlap with the purpose of the input message)? Might candidates be encouraged to copy (or reformulate) from the prompt instead of writing an original response?

W16	Prompt (instructions + input message) content: appropriate, culturally unbiased, not disturbing, suitable for a general-purpose test (i.e. not a specific purpose test)	2 – prompt is appropriate for international adult candidates, does not have any culturally biased content, will not emotionally disturb candidates; 1 – prompt might be somewhat inappropriate for a minor part of the candidates because of slight cultural bias and/or slight risk of disturbing; 0 – prompt is inappropriate because it deals with religion, violence, abuse, negative emotions (death, divorce, or other disturbing topics), directly discusses controversial political issues OR is culturally biased and will not be clear to candidates who are not familiar with the culture of English-speaking countries.	The writing prompt has been written for adult international candidates: Is the content of the prompt appropriate for an international audience? Might the content of the prompt affect performance of some of the candidates because of cultural bias? Is the prompt inappropriate because it is offensive or culturally insensitive? E.g. it deals with religion, violence, abuse, controversial political issues, might provoke a memory of negative events in a candidate's life?
OA	Overall acceptability of the prompt for inclusion in a test	2 – on the whole, the prompt can be accepted in its present form OR after minor revision; 1 – on the whole, the prompt requires major revision to be accepted; 0 – on the whole, the prompt should be rejected.	

Item evaluation scales: B1 listening tasks

Objectively-scored criteria		
	Evaluation criteria	**Rating scales**
L1	Text: max. 300 words	2 – text is 300 words max; 1 – text is up to 30 words over the limit AND is easy to reduce without any changes to the items; 0 – text is more than 30 words over the limit OR up to 30 words over the limit but cannot be reduced without making changes to the items.

L2	Text: lexis K1-K3 (1% of lexis can be proper names off frequency lists)	2 – all lexis is appropriate for the level (i.e. K1-K3); 1 – up to 4 words are above the stated norm for the level; 0 – more than four words are above the stated norm for the level.
L3	Topic: From the list of topics for B1 level	2 – topic stated is appropriate for the level and is accurately reflected in the text; 1 – topic stated is appropriate for the level BUT is not accurately reflected in the text; 0 – topic stated is not in the list for the level.
L4	Function: From the list of functions for B1 level	2 – function stated is appropriate for the level and is accurately reflected in the text; 1 – function stated is appropriate for the level BUT is not accurately reflected in the text; 0 – function stated is not in the list for the level.
L5	Items: 6 in total	2 – 6 items in the task; 1 – more than 6 items in the task; 0 – less than 6 items in the task.
L6	Items: either a set of notes or individual sentences	2 – all items are either individual sentences OR form a coherent set of notes; 1 – two items are included in one sentence OR one item consists of more than one sentence OR the set of notes is not fully coherent; 0 – several instances when two items are included in one sentence OR two or more items consist of more than one sentence OR the set of notes is totally incoherent.
L7	Stem: Max 10 words including the key	2 – each stem is max. 10 words including the key; 1 – one stem is 1-3 words over the limit AND is easy to reduce without much change to the stem; 0 – more than one stem is 1-3 words over the limit OR one or more stems is more than 3 words over the limit OR only one stem is up to 3 words over the limit but cannot be reduced without much change to the stem.
L8	Stem: lexis K1-K2	2 – all lexis is appropriate, i.e. K1-K2; 1 – up to one word for the six stems is above the stated norm for the level; 0 – more than one word for the six stems is above the stated norm for the level.
L9	Stem: grammar A1-A2	2 – all grammar is appropriate, i.e. A1-A2; 1 – one grammar structure is above the stated norm; 0 – more than one grammar structure is above the stated norm.
L10	Stem: a paraphrase, i.e. does not literally repeat what is heard in the text	2 – none of the stems literally repeats what is heard in the text but is a paraphrase; 1 – one stem (or a part of it) literally repeats what is heard in the text; 0 – more than one stem (or a part of it) literally repeats what is heard in the text.

L11	Response: lexis K1-K2 (except for proper names that are spelt out, there should be no more than 1 item of this kind per task)	2 – all lexis is appropriate, i.e.K1-K2; 1 – one word is above the stated norm OR a proper noun tested is not spelt out but is expected to be known to most candidates; 0 – more than one word is above the stated norm OR a proper noun tested is not spelt out and is not expected to be known to most candidates OR more than one proper noun is tested.	
L12	Spelling / grammar / punctuation: correct, including the text, items and the key	2 – all grammar, spelling and punctuation of the text, items and the key are correct; 1 – up to three grammar, spelling or punctuation errors; 0 – more than three grammar, spelling or punctuation errors.	
Subjectively-scored criteria			
	Evaluation criteria	**Rating scales**	**Guidance for reviewers**
L13	Text: A monologue (recorded instructions, lectures, presentations, public announcements, TV/radio programmes, short talks, news reports).	2 – text is a monologue in one of the specified genres 1 – text is a monologue but the genre, although largely appropriate, is not among the ones specified; 0 - text is not a monologue AND/OR the genre is not appropriate	Is the text a monologue? Is the genre of the text included in the list of genres from the specifications? If the genre is not mentioned in the specifications, is it still appropriate for the task?
L14	Text: sounds authentic according to the genre	2 – text sounds fully authentic according to the genre, i.e. one would expect to hear a similar sounding text in a real-life situation. 1text sounds mostly authentic according to the genre, while some minor parts do not; 0 – text sounds inauthentic according to the genre.	Does the text sound authentic according to the genre? Would you expect to hear a text like this in a real-life situation? If the text does not sound fully authentic, is it only parts of the text that sound inauthentic, or the whole text?
L15	Text: accessible to a B1 level test-taker	2 – the text will be accessible to B1-level test-takers as described in the CEFR; 1 – the text is mostly accessible to B1-level test-takers while some features might cause some difficulty or might be easier;	Would the text be accessible to B1-level test-takers, as described in the CEFR? Might the text be too difficult for B1-level test-takers because of the high density of information, absence of redundancies, syntactical complexity, etc.?

		0 – the text is not suitable for B1-level test-takers either because it is too difficult or too simple.	Might the text be insufficiently challenging for B1-level test-takers?
L16	Text: the content is appropriate, culturally unbiased, not disturbing	2 – text content is appropriate for international adult candidates, does not have any culturally biased content, will not emotionally disturb candidates; 1 – text content might be deemed inappropriate for a minority of the candidates because of slight cultural bias and/or being slightly disturbing; 0 – text content is inappropriate because it deals with religion, violence, abuse, negative emotions (death, divorce or other disturbing topics), directly discusses controversial political issues OR is culturally biased and will not be clear to candidates who are not familiar with the culture of English-speaking countries.	The listening task has been written for adult international candidates: Is the text content appropriate for an international audience? Might the text content affect the performance of some of the candidates because of cultural bias? Is the text content inappropriate because it is offensive or culturally insensitive? E.g. it deals with religion, violence, abuse, controversial political issues, might provoke a memory of negative events in a candidate's life?
L17	Text: suitable for testing, i.e. is NOT a parody, not silly, humorous, sarcastic, etc.	2 – text is suitable for testing purposes, i.e. is not a parody, not silly, humorous, sarcastic, or anything else that would be considered unsuitable in a testing situation; 1 - text is humorous BUT this will most probably not have any negative effect on the testing outcomes; 0 - input message is unsuitable for testing purposes because it is a parody, is silly humorous, sarcastic, or alternative material that would be considered unsuitable in a testing situation.	Is the text suitable for a testing situation? Does it contain any humour, sarcasm or any other connotations that might be unsuitable in a test? Will the testing outcomes be affected because of the inappropriate input message?
L18	Instruction: standard format is followed	2 - Instructions include all specified information (the speaker, the situation, guidance in how to fill the gaps); 1 – one piece of information is not fully presented or missing OR the instructions are redundant / somewhat awkwardly formulated;	

		0 – two or more pieces of information are missing from the instructions.	
L19	Items: test the ability to locate and record specific information from a monologue	2 – clear focus on candidate ability to locate and record specific information; 1 – one item does not (or loosely) test candidate ability to locate and record specific information OR tests specific information but is not clearly formulated 0 – two or more items do not test candidate ability to locate and record specific information OR are not clearly formulated	Is the ability to locate and record specific information tested in the task? Do all six items focus on testing candidate ability to locate and record specific information? Are there any items that focus on a different listening sub-skill?
L20	Items: do not test abilities unrelated to listening comprehension (e.g. maths, grammar, etc.)	2 - items do not test abilities unrelated to listening comprehension; 1 – one of the items tests an ability unrelated to listening comprehension e.g. maths, grammar, etc.; 0 – more than one item test abilities unrelated to listening comprehension.	
L21	Items: each item (except for proper names that are spelt out) has one or two pieces of information in the text that act as a distractor	2 – each item has at least one piece of information in the text that acts as a distractor; 1 – one item does not have any distractors in the text; 0 – two or more items do not have any distractors in the text.	Is there information in the text that acts as a distractor in the text? Does each item have at least one piece of information that acts as a distractor?
L22	Items: follow the order in the text	2 – all items follow the order of information as it appears in the text; 1 – one item does not follow the order of information as it appears in the text; 0 – more than one item does not follow the order of information as it appears in the text.	
L23	Items: The necessary information for different items is distributed	2 - the necessary information for different items is distributed across the whole text with no two pieces of	Is the necessary information for different items is distributed across the whole text with adequate distance

	across the whole text with no two pieces of information appearing too close to each other in the text	information appearing too close to each other in the text; 1 – the necessary information for 2 adjacent items appears very close together in the text; 0 - the necessary information for more than 2 items appears very close together in the text;	from one piece of information to the other for candidates to record their answers? Do any two pieces of information appear too close to each other, e.g. in the same line of the text?
L24	Stem: is clearly formulated in such a way that it restricts the number of possible correct answers	2 – each stem is clearly formulated and unambiguous so that candidates are sufficiently clear about what kind of information is required to fill in the gap; 1 – one of the items is not clearly formulated or is too vague, so that it would not be sufficiently clear to candidates what kind of information is required to fill in the gap; 0 – more than one item is not clearly formulated or is too vague, so that it would not be sufficiently clear to candidates what kind of information is required to fill in the gaps.	Is each stem formulated clearly? Is each stem unambiguous? Will the candidate be sufficiently clear about what kind of information is required to fill in each gap? Is there a clear single piece of information in the text to respond to each item? Can any item be answered with more than one piece of information from the text, all of which would be correct, according to the text?
L25	Response: requires max. 3 words or a number heard in the text	2 – each response requires max. 3 words or a number heard in the text; 1 – one response is 1-2 words over the limit OR is not heard in the text verbatim; 0 – one response is more than 2 words over the limit OR more than one response is 1-2 words over the limit OR more than one response is not heard in the text verbatim.	
L26	Response: All acceptable answers are included in the key	2 – all possible acceptable answers (for all items) are included in the key; 1 – one acceptable answer for one of the items is missing OR answers included are not present in the text OR (a) word(s) in one key overlap(s) with the stem.	Have all possible response versions (for all items) been included in the key?

		E.g. Stem: **The movie is showing at _____ pm.** Key: **9pm.** 0 – more than one acceptable answer for one item is missing AND/OR more than one item has (a) missing answer(s) AND/OR (a) word(s) in two or more keys overlap(s) with the stem.	
OA	Overall acceptability of the task for inclusion in a test	2 – on the whole, the task can be accepted in its present form OR after minor revision; 1 – on the whole, the task requires major revision to be accepted; 0 – on the whole, the task should be rejected.	Can the task be accepted as it is? Does the task require any revision? Is it just minor revision or major one? Is it at all possible to revise the task for it to be accepted, or should it be rejected, and a completely new task written?

Appendix 7: Tests of normality

A2 grammar items’ score totals	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Obj total Pre	.235	25	.001	.863	25	.003
Obj total Post	.309	25	.000	.591	25	.000
Subj total Pre	.248	25	.000	.887	25	.010
Subj total Post	.270	25	.000	.845	25	.001
Overall Pre	.158	25	.110	.930	25	.086
Overall Post	.233	25	.001	.754	25	.000

a. Lilliefors Significance Correction

C1 grammar items’ score totals	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Obj total Pre	.136	25	.200[*]	.919	25	.047
Obj total Post	.261	25	.000	.765	25	.000
Subj total Pre	.182	25	.033	.879	25	.007
Subj total Post	.209	25	.006	.920	25	.051
Overall Pre	.156	25	.121	.920	25	.050
Overall Post	.153	25	.133	.889	25	.011

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

B2 writing prompts’ score totals	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Obj total Pre	.230	25	.001	.843	25	.001
Obj total Post	.296	25	.000	.768	25	.000
Subj total Pre	.251	25	.000	.645	25	.000
Subj total Post	.176	25	.044	.893	25	.013

Overall Pre	.220	25	.003	.893	25	.013
Overall Post	.221	25	.003	.927	25	.075

a. Lilliefors Significance Correction

B1 listening tasks' score totals	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Obj total Pre	.121	25	.200[*]	.916	25	.041
Obj total Post	.201	25	.010	.928	25	.077
Subj total Pre	.142	25	.200[*]	.944	25	.179
Subj total Post	.232	25	.001	.916	25	.041
Overall Pre	.158	25	.108	.953	25	.290
Overall Post	.173	25	.051	.945	25	.193

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Appendix 8: Interview protocol

BEFORE THE INTERVIEW

1) *Introduce myself*
2) *Introduce my research:*

I'm conducting research into online item-writing training. In this interview, I will ask you to reflect on the item writing you have done recently – your responses will help me greatly in understanding the process of item writing and how the course may have helped you with this. *(FOR POST-COURSE INTERVIEWS: I would like to get feedback on the course for future runs).*

3) The interview will last 20-30 minutes (pre-course) / 30-40 minutes (post-course).
4) *Review the consent form:*

All personal information about you (e.g. your name and other information about you that can identify you) will be kept confidential and not shared with others. Findings from this study will be used for academic purposes to write a PhD thesis, make conference presentations and write journal articles. Your name will never be mentioned and any information that might identify you will not be revealed.

5) This interview is going to be recorded. Is it Ok with you?
6) Finally, if for any reason, you do not wish to answer any of the questions I ask, you may decline to do so.

Thank you for agreeing to take part in this interview!

PRE-COURSE INTERVIEW SCHEDULE: QUESTIONS AND PROMPTS

Main questions	**Follow-up questions**	**Reminder: per item type**
Can you tell me about your item-writing task? How did it go?	• How long did it take you to do the tasks? • Did you do them all in one go or did the item writing stretch over several days?	Grammar items? Writing prompt? Listening text? Listening task?
	• How did you approach writing the items? What influenced your decision to approach the task this way? • Did you use any resources / documents for writing the items?	Grammar items? Writing prompt? Listening text? Listening task?

	- Did you find it easy to write the items? - Did you have any difficulties? What kind of difficulties?	Grammar items? Writing prompt? Listening text? Listening task?
	- Do you feel there is any knowledge or there are any skills (you are lacking) that could have helped you write the items?	Grammar items? Writing prompt? Listening text? Listening task?
Is there anything else you would like to tell me about your item writing experience?		

POST-COURSE INTERVIEW SCHEDULE: QUESTIONS AND PROMPTS

Main questions	**Follow-up questions**	**Reminder: per item type**
Can you tell me about your item-writing task? How did it go?	- How long did it take you to do the tasks? - Did you do them all in one go or did the item writing stretch over several days?	Grammar items? Writing prompt? Listening text? Listening task?
	- How did you approach writing the items? What influenced your decision to approach the task this way? - Did you use any resources / documents / online tools for writing the items?	Grammar items? Writing prompt? Listening text? Listening task?
	- Did you find it easy to write the items? - Did you have any difficulties? What kind of difficulties?	Grammar items? Writing prompt? Listening text? Listening task?

Please tell me more about the item-writing course you took. Do you think the course has helped you in any way to write items?	• Do you feel more confident writing the items after you've taken the course? • Were there any particular aspects of the training course that have helped you in writing the items? • Is there anything particular not covered in the course and which would have helped you to write the items? • How ready to you feel to start doing item-writing work?	• How useful was the input material? (videos, ppt presentations, optional articles) • How useful were the course activities? (quizzes, group discussions, item-writing practice, reviewing items in groups)
Is there anything else you would like to tell me about your item-writing experience? About the item-writing course?		

TO SAY AFTER THE INTERVIEW

- Thanks for your time to participate in the interview. The information and opinions you have shared are very useful and will help a lot with the research study.
- I am going to conduct some more interviews with your colleagues in the next several weeks and then review the interview recordings. If I need to clarify some information you have provided, would it be OK if I approach you via email with some questions? It's OK to say 'no' if you would not like to.

If you have any questions after the interview, you can always contact me via my email o.rossi@lancaster.ac.uk

Thanks again for your time and your insights, it was a pleasure talking to you!

Appendix 9: Interview coding scheme

	PRE-TRAINING	POST-TRAINING
ITEM-WRITING SKILL DEVELOPMENT	Total: 420	Total: 423
Overall comments:	**Total: 145**	**Total: 156**
Perceived difficulty of item-writing	76	104
Attitude to item-writing	27	24
Use of specifications	16	20
Use of example items	26	8
Grammar item-specific comments:	**Total: 71**	**Total: 64**
Grammar_objective_topic&function	18	5
Grammar_objective_vocabulary frequency	8	11
Grammar_objective_word-limit	4	1
Grammar_subjective_construct_A2	8	6
Grammar_subjective_construct_C1	9	10
Grammar_subjective_stem	4	10
Grammar_subjective_key&distractors	12	18
Grammar_own grammar knowledge	8	3
Writing prompt-specific comments:	**Total: 53**	**Total: 63**
Writing_objective_grammar level	7	8
Writing_objective_vocabulary frequency	9	10
Writing_objective_word-limit	5	8
Writing_subjective_construct	2	7
Writing_subjective_input message	18	17
Writing_subjective_authenticity	6	10
Writing_subjective_instructions	6	3
Listening task-specific comments:	**Total: 151**	**Total: 140**
Listening_objective_topic	14	17
Listening_objective_vocabulary frequency	29	10

Listening_objective_word-limit	9	5
Listening_objective_grammar level	3	2
Listening_subjective_construct	2	4
Listening_subjective_input text genre	16	21
Listening_subjective_input text authenticity	20	28
Listening_subjective_items	35	28
Listening_subjective_in-text distractors	23	25
ROLE OF THE TRAINING		Total: 95
Training materials:		**Total: 43**
Materials in general	N/A	3
Materials to introduce the theory of item-writing	N/A	22
PPTs	N/A	10
Example items	N/A	7
Quizzes	N/A	1
Training activities:		**Total: 40**
Group discussions	N/A	3
Preparation for item-writing practice	N/A	3
Item-writing practice	N/A	12
Tutor feedback	N/A	11
Peer feedback	N/A	11
Course structure:		**Total: 8**
Course structure		8
Use of technology:		**Total: 4**
Lextutor	N/A	2
Wechat	N/A	2

Appendix 10: Feedback questionnaire 1

Your general impression of the course materials

Q1 On a scale from 0-10, how USEFUL do you feel the materials of modules 1 and 2 were?

Q2 Please elaborate on why you answered the previous question the way you did:

____

Q3 On a scale from 0-10, how INTERESTING do you feel the materials of modules 1 and 2 were?

Q4 Please elaborate on why you answered the previous question the way you did:

____

Q5 On a scale from 0-10, how USER-FRIENDLY do you feel the materials of modules 1 and 2 were?

Q6 Please elaborate on why you answered the previous question the way you did:

____

Q7 On a scale from 0-10, how would you rate the QUALITY of materials in modules 1 and 2?

Q8 Please elaborate on why you answered the previous question the way you did:

____

Your opinion about specific materials in Modules 1 and 2

Q9 Please indicate how useful you feel the following materials from modules 1 and 2 were in helping to develop your item-writing skills: (totally useless – useless – somewhat useless – somewhat useful – useful – extremely useful). Please explain.

- Power point presentation about the CEFR (module 1 task 1)
- Article about the CEFR and item writing by Davidson & Fulcher (module 1 task 2)
- Power point presentation 'What makes a good test item' (module 1 task 4)
- Quiz 'Dos and Don'ts of item writing' (module 1 task 4)
- Power point presentation about item specifications and QR checklists (module 1 task 5)
- Power point presentation about the Core Inventory document (module 2 task 1)
- Quiz on the Core Inventory document (module 2 task 1)
- Power point presentation on how to write multiple choice items (module 2 task 2)
- Ten weak multiple choice items for analysis and group discussion (module 2 task 2)
- Power point presentation about the construct of grammar items (module 2 task 3)

- Item-writing task (module 2 task 4)

Your suggestions for material improvement

Q10 If you have any suggestions on how the course materials could be improved to better help you develop your item-writing skills, please provide them in the space below:

____

Q11 What other (types of) materials can you think of that could help you develop your item-writing skills?

____

Q12 Finally, is there anything else you would like to share about the course?

____

Appendix 11: Feedback questionnaire 2

Your general impression of the course activities

Q1 On a scale from 0-10, how USEFUL do you feel the activities of modules 3 and 4 were?

Q2 Please elaborate on why you answered the previous question the way you did:

____

Q3 On a scale from 0-10, how INTERESTING do you feel the activities of modules 3 and 4 were?

Q4 Please elaborate on why you answered the previous question the way you did:

____

Q5 On a scale from 0-10, how USER-FRIENDLY do you feel the activities of modules 3 and 4 were?

Q6 Please elaborate on why you answered the previous question the way you did:

____

Your opinion about specific activities in modules 3 and 4

Q7 Please indicate how USEFUL you feel the following activities from modules 3 and 4 were in helping to develop your item-writing skills (totally useless – useless – somewhat useless – somewhat useful – useful – extremely useful). Please explain.

- Adjusting vocabulary difficulty of an authentic text to B1 level (module 3 task 1
- Reading and discussing and article on lexical competence by Meara (1996) (module 3 task 2)
- Analysing weak multiple matching vocabulary tasks (module 3 task 3) and writing prompts (module 4 task 3)
- Writing a multiple matching vocabulary task (module 3 task 4)
- Designing a range of speaking and writing prompts (module 4 tasks 1, 2, and 4)
- Doing item quality review in groups (module 3 task 4, module 4 tasks 1, 2, and 4)
- Receiving individual feedback on your work from the course tutors via email (module 3 task 1 and 4, module 4 task 2 and 4)
- Receiving group feedback on your work from the course tutor (task summaries for module 3 task 2 and 3; module 4 task 1 and 3)

Q8 Please indicate your preference for the modes of interaction used in activities of modules 3 and 4 from MOST PREFERRED (#1) to LEAST PREFERRED (#3). Drag and drop each statement to change its position in the list.

______ working individually

______ working in groups (e.g. discussing writing prompts, doing quality review)

______ a combination of both

Q9 Please elaborate on why you ranked the modes of interaction the way you did:

____

Your suggestions for activities improvement

Q10 If you have any suggestions on how the course activities could be improved to better help you develop your item-writing skills, please provide them in the space below:

____

Q11 What other (types of) activities can you think of that could help you develop your item-writing skills?

____

Q12 Finally, is there anything else you would like to share about the course?

____

Appendix 12: Feedback questionnaire 3

Your impression of the course structure: the course as a whole

Q1 Do you agree with the statement "The course was WELL-STRUCTURED"? Please indicate your agreement on a scale from 0-10

Q2 Please elaborate on why you answered the previous question the way you did:

____

Q3 On a scale from 0 to 10, how would you rate the CLARITY of the course structure?

Q4 Please elaborate on why you answered the previous question the way you did:

____

Q5 How appropriate was the FLEXIBILITY of the course structure?

○ The course structure was appropriately flexible

○ The course structure was appropriately inflexible

○ The course structure was not flexible enough

○ The course structure was too flexible

○ Other (please specify) ____

Q6 Please elaborate on why you answered the previous question the way you did:

____

Q7 How appropriate was the PACE of the course in terms of its EVENNESS?

○ The course pace was appropriately even

○ The course pace was appropriately uneven / varied

○ The course pace was too even

○ the course pace was too uneven

○ Other (please specify) ____

Q8 Please elaborate on why you answered the previous question the way you did:

____

Q9 How appropriate was the PACE of the course in terms of its SPEED?

○ The course pace was appropriately fast

○ The course pace was appropriately slow

○ The course pace was neither fast nor slow, which was appropriate

○ The course pace was too fast

○ The course pace was too slow

○ Other (please specify) ____

Q10 Please elaborate on why you answered the previous question the way you did:

____

Q11 If you have any suggestions on how the overall course structure could be improved to better help you develop your item-writing skills, please provide them in the space below:

____

Your opinion about the structure of a specific module: Module 6

Q12 Please indicate how USEFUL you feel the STRUCTURE of Module 6 'Listening' was in helping to develop your item-writing skills (totally useless – useless – somewhat useless – somewhat useful – useful – extremely useful). Please explain.

- Task **sequencing**: Task 1 (the construct of listening assessment + making changes to the pre-course listening text) - Task 2 (textmapping) - Task 3 (listening item writing)
- Use of **interactive activities** within the module (group discussions, giving feedback on items)

Q13 Please indicate how APPROPRIATE you feel the STRUCTURE of Module 6 'Listening' was in helping to develop your item writing skills (totally inappropriate – inappropriate – somewhat inappropriate – somewhat appropriate– appropriate – fully appropriate). Please explain.

- Module 6 **flexibility**
- Module 6 **pace** in terms of its **evenness and speed**

Q14 If you have any suggestions on how the Module 6 'Listening' structure could be improved to better help you develop your item-writing skills, please provide them in the space below:

____

Your general impression of the technology used on the course (Edmodo, Wechat, email): the course as a whole

Q15 On a scale from 0-10, how USEFUL do you feel the technology was?

Q16 Please elaborate on why you answered the previous question the way you did:

____

Q17 On a scale from 0-10, how SUPPORTIVE do you feel the use of technology was in delivering course aims?

Q18 Please elaborate on why you answered the previous question the way you did:

____

Q19 On a scale from 0-10, how USER-FRIENDLY do you feel the technology was?

Q20 Please elaborate on why you answered the previous question the way you did:

____

Your opinion about the use of technology in a specific module: Module 6

Q21 Please indicate how USEFUL you feel the technology was in helping to develop your item-writing skills (totally useless – useless – somewhat useless – somewhat useful – useful – extremely useful). Please explain.

- Using Edmodo to introduce the module aims and activities
- Using Edmodo library to store module materials
- Using Wechat groups to discuss pre-course listening texts and do the textmapping activity
- Using Wechat groups to give feedback on each other's listening items (task 3)
- Using email to submit listening items (task 3)
- Using email to receive individual feedback from the course tutors (task 3)
- Using Edmodo to access task summaries

Your suggestions on how to improve the use of technology on the course

Q22 If you have any suggestions on how the use of technology could be improved to better help you develop your item-writing skills, please provide them in the space below:

____

Q23 What other (types of) technology can you think of that could help you develop your item-writing skills and be embedded in an item writer course?

____

Q24 Finally, is there anything else you would like to share about the course?

____

Appendix 13: Final feedback questionnaire

This is the final item-writing course feedback questionnaire. It will ask you for your views on the course AS A WHOLE, not about individual modules. The questionnaire is very brief and will take 2-3 minutes of your time at most. You will not have to write anything - just choose a response. Thank you for your valuable insights into item-writing training!

Course materials overall

Q1 On a scale from 0-10, how USEFUL do you feel the course materials were?

Q2 On a scale from 0-10, how INTERESTING do you feel the course materials were?

Q3 On a scale from 0-10, how USER-FRIENDLY do you feel the course materials were?

Q4 On a scale from 0-10, how would you rate the QUALITY of course materials?

Q5 If you have any suggestions on how course materials could be improved, please provide them in the space below:

____

Course activities overall

Q6 On a scale from 0-10, how USEFUL do you feel the course activities were?

Q7 On a scale from 0-10, how INTERESTING do you feel the course activities were?

Q8 On a scale from 0-10, how USER-FRIENDLY do you feel the course activities were?

Q9 If you have any suggestions on how course activities could be improved, please provide them in the space below:

____

Course structure overall

Q10 Do you agree with the statement "The course was WELL-STRUCTURED"? Please indicate your agreement on a scale from 0-10

Q11 On a scale from 0 to 10, how would you rate the CLARITY of the course structure?

Q12 How appropriate was the FLEXIBILITY of the course structure?

○ appropriately flexible

○ appropriately inflexible

○ not flexible enough

○ too flexible

Q13 How appropriate was the PACE of the course in terms of its EVENNESS?

○ appropriately even

○ appropriately uneven / varied

○ too even

○ too uneven

Q14 How appropriate was the PACE of the course in terms of its SPEED?

○ appropriately fast

○ appropriately slow

○ neither fast nor slow, which was appropriate

○ too fast

○ too slow

Q15 If you have any suggestions on how the overall course structure could be improved, please provide them in the space below:

____

Use of technology overall

Q16 On a scale from 0-10, how USEFUL do you feel the technology was throughout the course?

Q17 On a scale from 0-10, how SUPPORTIVE do you feel the use of technology was in delivering the course aims?

Q18 On a scale from 0-10, how USER-FRIENDLY do you feel the technology was throughout the course?

Q19 If you have any suggestions on how the use of technology could be improved, please provide them in the space below:

____

Finally,

Q20 ... is there anything else you would like to share about the course?

____

Thank you for completing the questionnaire!

Appendix 14: Band score frequencies

				Band 0	Band 1	Band 2
Grammar items	A2	Objectively-scored criteria	pre	10.4%	4.8%	84.8%
			post	1.6%	4%	94.4%
	C1		pre	12.8%	6.8%	80.4%
			post	2.4%	5.2%	92.4%
	A2	Subjectively-scored criteria	pre	8%	19%	73%
			post	3.5%	18.5%	78%
	C1		pre	5%	21%	74%
			post	2%	20%	78%
Writing B2 prompts		Objectively-scored criteria	pre	4%	16.8%	79.2%
			post	2.4%	12%	85.6%
		Subjectively-scored criteria	pre	2.5%	14.2%	83.3%
			post	1.8%	10.2%	88%
Listening B1 tasks		Objectively-scored criteria	pre	13.4%	19.3%	67.3%
			post	6%	15.7%	78.3%
		Subjectively-scored criteria	pre	6.3%	19.1%	74.6%
			post	3.2%	19.4%	77.4%

Appendix 15: Empirical studies of item-writing training effectiveness

Study / field / country	Study design	Participants	Item type	Intervention type	Data type
Abdulghani et al. (2015) **Medicine (respiratory, cardiovascular, and renal)** **Saudi Arabia**	Pretest-posttest, no control group	25 newly-joined faculty members	MCQs	Two full-day face-to-face workshops. Day 1: theoretical background, item flaws, revision of past MCQs. Day 2: writing MCQs in groups of 3-4 participants using a checklist.	Tests produced in the year before and after the training: Item difficulty and discrimination, non-functioning distractors, students' performance analysed for the test as a whole, irrespective of whether the items were produced by participants or non-participants.
Dellinges & Curtis (2017) **Dentistry** **USA**	Pretest-posttest, experimental and control groups	Dental school faculty with previous item-writing experience (12 in the experimental and 12 in the control group),	MCQs	1-hr face-to-face session: 30-min PowerPoint presentation on ways to increase MCQ quality + discussion of poorly constructed and improved MCQ items.	Two versions of 6 MCQs per participant: produced before the training and then improved. Evaluated by 2 judges (blinded) against a 7-criterion 2-band rating scale. Scores from 2 judges averaged.
Gupta et al. (2020) **Medicine (various)** **India**	Pretest-pottest, no control group	28 medical college faculty with 3-30 years of teaching experience (M=10)	MCQs	3-hr face-to-face session: input in producing MCQs according to official guidelines for medical faculty	1. MCQs produced by participants before and after the training. Analysed for 16 MCQ flaws. Number of judges and whether they were blinded unknown. 2. 50 pretest and 50 posttest items used in a live test. Responses analysed for item

					difficulty, discrimination, and non-functioning distractors.
Hamamoto Filho & Bicudo (2020) **Medicine (various)** **Brazil**	Pretest-posttest, no control group	Medical school faculty, number and experience unknown	MCQs	Feedback on items produced by the faculty as a whole in the previous year: quality of items, changes made by the review panel, students' performance, item performance (difficulty and discrimination)	Items submitted by the faculty for inclusion in two tests (one in the year before and one after the feedback). Evaluated against a 7-criterion 2-band rating scale. Number of judges and whether they were blinded unknown.
Iramaneerat (2012) **Medicine (various)** **Thailand**	Pretest-posttest, experimental and control groups	Medical school faculty with previous item-writing experience. Experimental group: 68 in 1st workshop, 51 in 2nd & 3rd workshop. Control group: unknown.	MCQs	3 face-to-face workshops: 1) 3-hr session on MCQ item development and common flaws; 2 and 3) 2-hr input on classical item analysis and how to use it to improve item quality.	1.Items produced in the year before and after the workshop. Item difficulty and discrimination analysed separately for participants and non-participants. 2. Quantitative responses to training satisfaction questionnaires administered after each workshop
Naeem et al. (2012) **Medicine (various)** **Pakistan**	Pretest-midtest-posttest, no control group	51 faculty members with previous item-writing experience	MCQs, short-answer questions, Objective Structured Clinical Examination checklists	One-week full-time face-to-face training: presentations, item-writing practice, peer- and trainer-feedback.	Three versions of 3 items per participant (one of each type): produced before the training and then improved in two steps following the trainer- and peer-feedback. Evaluated by one judge (not blinded) against a 21-criterion (MCQs) / 16-criterion (SAQs) / 21-criterion (OSCEs) 2-band rating scale.

Scott et al. (2019) **Medicine (emergency)** **USA**	Pretest-posttest, no control group	16 students and resident volunteers inexperienced in item-writing	MCQs	30-min PowerPoint presentation with voice-over, watched online by all participants together on a conference call, followed with 10-min Q&A session.	3 MCQs produced by each participant before the training and 3 new MCQs produced immediately after the training. Evaluated by two judges (blinded) against a 7-criterion 2-band rating scale. Score discrepancies adjudicated by 3rd judge.
Tricio et al. (2018) **Dentistry** **Chile**	Pretest-posttest, no control group	Medical school faculty, number and experience unknown; 81% attended at least one training workshop	MCQs	Several workshops (exact number, length, and content unknown), a detailed item construction and blueprint guide, personalised guidance to improve items	1359 items produced in the year before and 1596 items produced in the year after the training. Evaluated against a 21-criterion 2-band rating scale. The number of judges and whether they were blinded unknown.
Yurdakul et al. (2020) **Mathematics** **Turkey**	Posttest only, no control group	100 school teachers with previous item-writing experience	MCQs, T/F, open-ended questions	Two full-day face-to-face workshops in producing higher-order thinking skills maths items. Day 1 - input; day 2 – item-writing practice.	1.Items produced during the training evaluated for the level of cognitive demand on a 4-level scale by 2 judges. 2.Quantitative and qualitative responses to participant feedback questionnaires.