What makes item-writing training useful? Learning from trainee perceptions on an online item-writing course for language testing

Olena Rossi[1], Tineke Brunfaut[2] & Luke Harding[2]

[1]Independent Language Assessment Consultant, Italy
[2]Lancaster University, UK

Despite general recognition that, in human-led item writing, training is key to producing good-quality tests, there is little empirical research on what constitutes good practice in item-writing training for language test development. This has led to a lack of evidence-based guidance for those who (plan to) organise item-writing training. To help address this research gap, this study explored the perceptions of 25 novice item writers on the usefulness of a three-month, online induction item-writing training course. Views were collected via four feedback questionnaires administered at fixed points throughout the course and in semi-structured interviews conducted on course completion. Findings showed that participants particularly valued a clear bite-size course structure, extensive item-writing practice, timely and detailed tutor feedback, and regular opportunities for peer collaboration. Combining language testing theory with item-writing practice was also viewed as beneficial for learning. Participants held mixed views, however, on the platforms used for course delivery. Based on the findings, practical recommendations are proposed for how training for human-led item-writing can be usefully structured and delivered.

Keywords: Item writing, item writers, item writer training, online training

Email address for correspondence: olena.rossi@itemwriting.co

Introduction

Item writers are those people who produce test items, ideally following item writing specifications. As item writers effectively create test content, their work is crucial to establishing test validity – a fundamental consideration in assessment (Bachman & Palmer, 2010). Poorly constructed items, or items that do not target intended constructs effectively, directly weaken inferences based on test scores. In practice, however, many item writers receive little or no formal training and have to learn to write items by trial and error (Alderson, 2010). Language teachers in various educational contexts, for example, are often expected to develop language assessments with limited to no item-writing training or experience (Green, 2016). Additionally, while many large-scale test organisations provide in-house training to their item writers, little information is publicly available on how this training is organised. Where accounts of item-writing training exist (e.g., Ingham, 2008; de Jong, 2008), authors do not typically report data-driven evaluations of the effectiveness of training or perceptions of its usefulness.

Language testing textbooks provide practical recommendations for producing good-quality items (e.g., Brown & Abeywickrama, 2018; Hughes & Hughes, 2020) but contain little guidance on how people can be trained to write those items. The field of educational measurement provides a wider literature on this topic (e.g., Haladyna & Rodriguez, 2013; Welch, 2006), but training recommendations are generally derived from authors' intuitions and impressions rather than empirical evaluations. There is, therefore, a need for evidence-based approaches to determine the extent to which trainees find item-writing training useful, and to identify which elements of item-writing training are particularly beneficial. Developing understanding of best-practice in item-writing training is of direct practical benefit to instructors/facilitators delivering item-writing training.

In this study, we address this gap with empirical data, investigating the perceptions and attitudes of English as a Foreign Language (EFL) teachers with no prior item writing experience during a three-month, online item-writing training programme. The effectiveness of this training course from a cognitive, skills development perspective, i.e. item-quality improvement, has already been reported in detail in

Rossi (2021) and Rossi and Brunfaut (2021). It was found that the item-writing training course resulted in statistically significant improvements in item quality, particularly for grammar and listening tasks. The gains were most evident on those aspects of items that can be objectively evaluated (e.g., reflecting the specified lexical frequency bands; topic and language functions from the specified list; no literal overlap between text input and item stem) and were observed across different trainee profiles, suggesting that the training led to measurable skill development.

In the current article, we focus on effectiveness from an affective perspective, i.e. trainees' emotional and attitudinal responses to the course - such as motivation, confidence, engagement, and perceived usefulness. This focus aligns with the constructivist framework (Mayes, 2001; Steffe & Gale, 1995) which underpinned the present course's design (see Context section below) and which emphasises active learner participation, collaborative knowledge-building, and reflection. These elements of learning are not solely cognitive but also deeply affective, as learners' emotions, beliefs, and attitudes can significantly influence how they engage with content and with peers (Mason, 2001). Investigating affective dimensions can thus help illuminate how item-writing trainees experience a course and how those experiences shape their learning. Our analysis of the data in this study provides insights for developing theory about good practice in item-writing training programmes, with implications for those conducting training in other contexts.

Literature review

Item writers' contribution to ensuring test validity is rarely made explicit in the language testing literature (Green & Hawkey, 2011; Rossi & Brunfaut, 2019). Shin (2022) lamented that item writers are often viewed as quasi-professionals, as compared to test designers, raters, rater trainers, and data analysts. Indeed, rater training has been the focus of much research (e.g., Knoch et al., 2016), while item-writing training remains relatively underexplored.

In many contexts, language teachers or those with teaching experience are often called to act as item writers in large-scale test development projects at institutional, regional or national levels (Baker & Riches, 2018; Brunfaut & Harding, 2018; Kremmel et al., 2018). Multiple studies have found, however, that teachers are often not prepared for

writing language test items due to inadequate training provision (Baker & Riches, 2018; Lam, 2015), and that they tend to rely on "test as you were tested" approaches (Tsagari & Vogt, 2017, p.54), focusing on limited item types and uncritical cloning of existing materials (Villa Larenas & Brunfaut, 2023). Lack of item-writing training was also documented by Alderson (2010), who surveyed organisations producing Aviation English tests and found that half did not provide training for their item writers. Several other sources (e.g., Osterlindt, 1998; Spaan, 2007) refer to instances where experienced item writers are expected to serve as mentors to novices, with development expected to happen 'on the job'; however, the specific training approach that might happen on the job is rarely documented.

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, or idiosyncratic items. For this reason, Downing (2006) argued that it is essential for anyone with responsibility for writing test items to be formally trained. The necessity of item-writing training has been confirmed, for example, through empirical studies in educational measurement for medical science. For instance, Jozefowicz et al. (2002) found a significant difference in quality ratings of medical examination multiple-choice questions (MCQs) written by faculty trained in item writing compared with faculty without such training.

While the observation that training improves item quality now seems well established, few publications provide practical recommendations for how to organise item-writing training. Downing (2006) advocated hands-on training workshops structured as an "instruction – practice – feedback – reinforcement loop" (p.11). Welch (2006, p.309) proposed an agenda for a face-to-face workshop on producing prompts for performance assessment:

(1) discussion on the purpose and audience of the assessment
(2) presentation of the test specifications and test development process
(3) general guidelines for prompt writing
(4) presentation of the prompt templates or "item shells"
(5) presentation and discussion of successful and unsuccessful prompts
(6) trainees generate topics for consideration followed by a discussion

(7) trainees create prompts from the approved topics.

This training outline is largely reiterated in the item-writing training schedule proposed by Haladyna and Rodriguez (2013). Importantly, the above-mentioned training schedules stress the necessity for item-writing practice and discussions of items within groups 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" (Haladyna & Rodriguez, 2013, p.23).

In the field of language testing, item-writing training is typically discussed in the context of prominent testing bodies who have their own training approaches and procedures. For example, Ingham (2008) and de Jong (2008) described how item-writing training was conducted (at the time of publication) at Cambridge Assessment and Pearson Education, respectively. A standard training weekend that served as an induction 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; (2) two-hour sessions on the techniques of writing particular item types, including group activities drawing on the ideas and experience of the participants; (3) an overview of writing for particular skills papers; and (4) text selection and adaptation (Ingham, 2008, pp.6-7). Item-writing training at Pearson Education, as explained by de Jong (2008), comprised a one-day face-to-face workshop covering: an introduction to the Common European Framework of Reference (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. Al-Lawati's (2014) study, investigating the use of the Pearson Test of English General specifications and item writing guidelines for producing reading test items, provided some rare insights into item-writers' own perceptions of their training needs. The interviewed item writers recommended that training should include "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 some language teacher education contexts, a shift towards greater attention to language assessment literacy (LAL) has spurred an increased focus on hands-on assessment practice (e.g., Giraldo & Murcia, 2019). This is because LAL research has

largely confirmed that teachers primarily value practice-orientated training. For example, Kremmel et al. (2018) found that teachers involved in item-writing training for the Austrian national school-leaving language examinations felt that the most useful type of activity mimicked the item writing process: "drawing up tasks and revising them after receiving feedback" (p.186). Importantly, the newly gained item-writing skills were seen by the teachers as applicable to classroom testing. This perspective was also shared by teachers in Luxembourg following item-writing training for the development of a national exam (see Harding & Brunfaut, 2020). Harsch et al. (2021) found that teachers receiving training in language assessment preferred learning about practical aspects of assessment, while teachers in Yan and Fan (2021) reported having learnt the most "from the discussion and revisions of the items" (p.231).

Results of teacher LAL studies also seem to confirm the positive effect of collaboration on learning about item writing (Baker & Riches, 2018; Cui et al., 2022). In Harsch et al.'s (2021) project, teachers "suggested working groups (WGs) to develop tasks collaboratively and give feedback to each other" (p.322). Ho and Yan (2021) studied an essay-prompt-writing training course at a US university. The semester-long training involved collaborative prompt development through the process of peer-feedback and revisions. The main finding was that item-writing collaboration played an important role in improving LAL overall.

While it seems clear from previous studies that hands-on item development and collaborative discussion are valued activities in LAL development more generally, there remain open questions concerning what features of more dedicated item writer training programmes are valued by participants. First, it is not clear what types of materials are preferred by trainees. Second, trainees' perceptions of course structure in item writer training for language test development – the sequencing and timing of activities – have been underexplored. And finally, given the prevalence of online training delivery (e.g. Borg, 2021), it is important to establish whether the preferred training activities identified in the preceding literature translate into online delivery, and – more broadly – what the general perceptions are concerning online delivery of item writer training.

To address these issues, we conducted a detailed evaluation of an online item-writing training course offered to EFL teachers aspiring to become item writers, with the present article focusing on effectiveness from an affective perspective (as mentioned, effectiveness from an item quality improvement perspective is reported in Rossi [2021] and Rossi and Brunfaut [2021]). We therefore investigated trainees' perceptions of and attitudes towards different elements of an item-writing training course. We sought to understand the trainees' perceptions and attitudes, and also to track whether these shifted over the course of the item-writing training course.

Tracking how trainees' perceptions and attitudes change over time can offer valuable insights into the learning process and inform training design. Previous research suggests that changes in attitudes may signal increased engagement with course content, growing confidence, or shifts in beliefs about key concepts. For example, a longitudinal study by Lowell and McNeill (2023) found that science teachers' instructional beliefs and self-efficacy developed at different rates over a two-year professional development programme. While instructional beliefs improved early on, self-efficacy increased more gradually, highlighting the need for sustained opportunities to practise and apply new knowledge. Huang et al. (2022) found that pharmacy students' attitudes towards professionalism significantly improved after a five-week experiential course, demonstrating that even relatively short interventions can influence professional dispositions.

In the context of item-writing training, where confidence and perceived relevance are especially important, changes in perceptions over time may reflect both cognitive and affective learning. Karthikeyan et al. (2019), in a scoping review of item-writing training in medical education, highlighted a gap in understanding how to effectively engage participants in item-writing training. This suggests a need to investigate not only item quality, but also how trainees perceive and respond to training processes - particularly as these perceptions may influence longer-term engagement and learning. Monitoring shifts in attitudes can help identify points of motivation or frustration, which can in turn guide improvements in task sequencing, scaffolding, and feedback mechanisms.

Examining changes in trainees' perceptions also supports a constructivist view of learning (Mayes, 2001), where understanding develops iteratively through experience

and reflection. From this perspective, attitudinal change is not simply a side effect of instruction but an integral part of professional development. Documenting these shifts allows course designers to evaluate not just what trainees have learned, but how they have responded to the training process over time - providing a more complete picture of the course's impact.

Two research questions were therefore posed:

RQ1: What were trainees' perceptions of and attitudes towards an item-writing training course?

RQ2: How did trainees' perceptions and attitudes develop from the beginning to the end of the course?

Context: The online item-writing training course

The online item-writing training course that formed the focus of this study was developed by the first author between 2016 and 2018 following a commission by the British Council's East-Asia Assessment Solutions Team. This author was also the lead tutor for the three training cohorts of the course. The trainees were 25 experienced EFL teachers who, at the time of training, worked as English language examiners for the British Council in locations across China. The British Council provided examiners with opportunities to develop their language assessment-related skills with potential to make future contributions to assessment projects in the East-Asia region; the item-writing training course formed one of these opportunities. The present study draws on data from training Cohort 3, the latest cohort at the time of conducting the research. The training course had been developed from scratch alongside Cohort 1 and the research methodology for this study and for Rossi (2021) and Rossi and Brunfaut (2021) had been piloted with 10 trainees during Cohort 2.

The course was 12 weeks long in total and consisted of six modules (see Table 1). Each module ran over two weeks and focused on a specific element of language assessment. The first Module offered a general introduction, covering broad topics such as writing items against specifications, item writing with reference to the CEFR, qualities of good items, and using checklists for item quality review. Modules 2 and 3 then focused on

grammar and vocabulary, respectively, which were purposively scheduled early in the course and allocated two weeks each to also build foundational skills in producing good-quality discrete-point item types (e.g., multiple-choice, true/false, matching). They were seen as a good starting point for trainees before moving onto skills-based item writing. Module 4 (productive skills) covered writing and speaking in a single two-week block, as it focused solely on prompt creation and did not include rubric development, given the course's focus on item writing rather than rating. Modules 5 and 6, covering reading and listening, respectively, were placed at the end of the course because item writing for these skills is arguably more complex as it includes text selection and adaptation as well as item construction. It was anticipated that, by that point, trainees would be sufficiently prepared for these more cognitively demanding item-writing tasks.

A four-to-five-hour time commitment was expected from participants weekly. In the first week of each module, participants were introduced to the focal item-writing topic, 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 two, participants wrote their own items according to specifications, peer-reviewed the items in their groups, and submitted the revised items to the course tutors for individual feedback. The course syllabus is available through the Open Science Framework (OSF): https://osf.io/jsg3e/?view_only=73b49730049440 51a65bdb6b13cfca96.

Table 1. Online induction item-writing training course structure

Module	Topic
1	Introduction to item writing
2	Writing grammar items
3	Writing vocabulary items
4	Writing productive skills tasks
5	Writing reading tasks
6	Writing listening tasks

The scope of the training was based on Fulcher's (2012) LAL definition, which includes both theoretical knowledge of language testing principles and practical ability in test development: "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).

Consequently, the main course objectives were to: (1) equip trainees with knowledge of language testing principles relevant to item writing, and (2) develop their practical ability to produce language test items according to specifications.

The course was taught by two tutors (the first author as lead tutor together with another British Council tutor). The course had a modular format and was delivered online asynchronously through Edmodo, a free Learning Management System (LMS). Within Edmodo, 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 groups of four-to-six people for group discussions and to give feedback on each other's items. WeChat (a Chinese social media messaging app) was used to facilitate online group discussions: participants normally communicated in writing by posting their messages in the group discussion space. Course assignments were submitted via email.

The specifications used by trainees to produce items during the course were not drawn from any particular test but developed for the purposes of the training, with large-scale general English language proficiency testing in mind. The specifications were developed by following the principles of the socio-cognitive framework for test development (O'Sullivan & Weir, 2011), which primarily guides British Council testing work; those same principles would likely inform test projects the trainees might ultimately write items for. Examples of specifications used during the training are available from the OSF: https://osf.io/jsg3e/?view_only=73b49730049440 51a65bdb6b13cfca96.

Mayes' (2001) three-stage constructivist framework for online course design informed the training approach. Accordingly, each module of the course was structured in a similar way, consisting of three stages (see Table 2). A detailed example of Module Six of the course – dedicated to producing listening test tasks – is described in Rossi and Brunfaut (2021).

Table 2. Item-writing training course: Module structure

Training stages	Mayes (2001)	Online item-writing training course
1	Conceptualisation: learners come to an initial understanding of a concept under review	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	Construction: "an activity in which the new understanding is brought to bear on a problem" (p.19)	Collaborative activities aimed at applying 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	Consolidation: results in full integration of the new understanding with the learners' general framework of knowledge	Item-writing practice (e.g., producing speaking prompts against a set of specifications) followed with peer-feedback in small groups.

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 (Steffe & Gale, 1995). Therefore, a large part of the training was dedicated to item-writing practice. Constructivism advocates learner co-operation whereby peers help each other in constructing their own knowledge (Mason, 2001). Thus, the course included online discussions of language testing concepts, group analyses of test items, and collaborative peer-feedback on items to operationalize the constructivist idea of learner co-operation in an online environment. The tutors' roles were conceptualized as item-writing experts and facilitators who introduced trainees to activities, guided them through the item-writing process, clarified uncertainties, and provided expert feedback on their items.

Methods

Item-writing trainees

Twenty-five participants self-selected for the item-writing training course from among a pool of EFL teachers working at the time as speaking/writing examiners for the British Council China. Nineteen were male, six female. Their age range was 29-60 years' old. Twenty-three were English-L1 speakers, with one Dutch-L1 and one Polish-L1 speaker. All held university degrees (BA-100%, MA-64%), an ESL/EFL qualification (CELTA or equivalent), and at least three years' experience teaching EFL

(M=8.5 years). Although everyone had practical classroom assessment experience, the teachers reported no previous training in test item writing.

Instruments and data collection

To inform answers to the RQs, a course feedback study was designed. For RQ1, which investigated trainees' perceptions of and attitudes towards the item-writing training course, three questionnaires and an interview protocol were developed.

The design of the feedback questionnaires followed recommendations for evaluation in training and human resource development (Newby, 1992; Phillips, 1991; Rae, 1999). Each questionnaire addressed a specific area for evaluation: training materials (FQ1), training activities (FQ2), course structure (FQ3), and use of technology (also FQ3). The areas for evaluation were adapted from Phillips (1991, p.161), with 'use of technology' added as recommended in Rae (1999) due to the course's online mode and the various technologies (Edmodo, WeChat, email) used to support its delivery.

Each questionnaire contained between 12 and 24 questions (depending on the range of training areas covered). Questions were a combination of closed and open types. For closed questions, responses were provided on a Likert-style 0-10 scale (e.g., "On a scale from 0-10, how USEFUL do you feel the materials of modules 1 and 2 were?"). This scale-range was adopted to encourage trainees to make fine-grained distinctions in their course evaluations. Each questionnaire started with questions targeting general perceptions about the specific aspect of the course in focus (e.g., training materials), and then proceeded to questions about specific materials, activities, etc. used on the course. Each questionnaire was administered at a different point during the training course: FQ1 after Module 2, FQ2 after Module 4, and FQ3 after Module 6.

The decision to divide feedback across FQ1–3 and to spread the questionnaires out over time was guided by practical and pedagogical considerations. Each questionnaire was intentionally designed to elicit rich, specific feedback on a particular aspect; however, repeating all aspects and associated questions at every time point would have placed an excessive burden on participants, who were already engaged in an intensive training course alongside their full-time jobs. Feedback areas were therefore distributed across three questionnaires and administered based on their relevance to

the course timeline. For example, 'training materials' were assessed in FQ1 because participants could reasonably evaluate some materials already after completing the first two modules. In contrast, 'course structure' was only included in FQ3, when participants had experienced the full course sequence. Additionally, giving participants feedback opportunities from early on and throughout the course, rather than only at the end, aligns with constructivist approaches and pedagogically signals recognition of the student voice.

The questionnaires were completed anonymously; 19 participants submitted FQ1, 22 FQ2, and 22 FQ3. The questionnaires are available from https://osf.io/jsg3e/?view_only=73b4973004944051a65bdb6b13cfca96.

In terms of the interviews for RQ1, after course completion semi-structured interviews were conducted by the first author using WeChat call. The main question asked was broad and open-ended to allow interviewees freedom in responses (Rubin & Rubin, 2005): "Please tell me about the item-writing course you took". Follow-up questions prompted participants to elaborate on course aspects that were beneficial for developing their item-writing ability and to suggest ideas for course improvement. Nineteen participants volunteered for the interviews, which lasted 13–30 minutes and were audio-recorded and transcribed. Written-style conventions were used for transcription (Kvale & Brinkmann, 2009).

With respect to RQ2, in order to detect any changes in trainees' attitudes over time, a fourth feedback questionnaire (Final FQ) was developed. It repeated the main questions (20 in total) from all three preceding questionnaires but with fewer open questions made mandatory. This Final FQ was administered upon course completion (prior to the interviews), with 19 participants completing it (anonymously). The Final FQ is also available from https://osf.io/jsg3e/?view_only=73b4973004944051a65bdb6b13cfca96.

Data analyses

Descriptive statistics were calculated for all closed questionnaire items. To explore changes in participants' attitudes over time, responses to FQ1-3 were compared with the relevant sections of the Final FQ. Inferential statistics were not run due to the small

dataset. Responses to open-ended questions were coded by the first author for 'key themes' raised by the majority of participants. These themes were identified through a process of reviewing the responses multiple times, finding key words or phrases to summarise each response, establishing themes and evaluating these in accordance with the research questions (Newby, 1992). Ten percent of the open-ended responses were double-coded by another researcher (a specialist in language testing), with a very high level of exact coding agreement obtained (93%), suggesting trustworthy first coding of the remainder of the dataset. Where relevant, comparisons were also made between participants' open-ended answers to FQ1-3 and the Final FQ.

The interview transcripts were double-checked for accuracy and then coded thematically (Braun & Clarke, 2012) in Atlas.ti. A combination of deductive and inductive coding was used. Initial codes were based on the main course components targeted in the research questions and feedback instruments (e.g., training materials, activities, structure, technology), but additional themes were allowed to emerge from the data through iterative reading and re-coding. Ten percent of the data was double-coded, demonstrating 85% coder agreement. The two coders then discussed the coding differences to reach agreement.

The final stage of qualitative analysis involved triangulation of the FQ and interview findings. Eight themes relevant to this study were identified that spanned both sets of analyses: 1) 'item-writing theory' refers to participants' views on the usefulness and accessibility of the theoretical input provided in the course; 2) 'example items' captures reactions to the use and perceived value of sample test items - both strong and weak - as learning tools; 3) 'item-writing practice' encompasses participants' experiences with producing items themselves and learning through hands-on application; 4) 'tutor- and peer-feedback' concerns views on any guidance and critique received during the course, both from instructors and fellow trainees; 5) 'modes of interaction' refers to preferences and reflections on individual versus collaborative learning tasks; 6) 'course structure' reflects evaluations of the overall pacing, sequencing, and internal consistency of the course design; 7) 'use of technology' summarises participant feedback on the digital tools used to deliver and facilitate the training; and 8) 'suggestions for course improvement' captures ideas and recommendations offered by trainees for enhancing future iterations of the course.

Results

First, we present quantitative findings from the feedback questionnaires. Then, we summarise qualitative findings from the qualitative questionnaire and interview data.

Participants' quantitative evaluations of the item-writing course

Table 3 provides descriptive statistics (medians and interquartile range [IQR]) for questionnaire items that required a scale response, comparing within-course feedback questionnaires (FQs 1-3) with corresponding items on the final questionnaire (Final FQ).

As shown in Table 3, participants had positive impressions of the training materials, with usefulness, interest, user-friendliness, and quality all receiving median ratings of 7 or 8 with increases of one point for usefulness and interest. Notably, interquartile ranges for responses on the final questionnaire were smaller compared to those provided in FQ1 for all items except interest, indicating that as the course progressed, participants' high ratings of the training materials generally became more uniform. Training activities were also rated highly in terms of usefulness, interest and user-friendliness, and interquartile ranges again showed that participants' views converged slightly between the administration of FQ2 and the final questionnaire. Course structure (well-structured? clear?) received strong appraisals across both questionnaires, though with slightly more spread among participants' ratings than training material and training activities. The lowest ratings (though with an IQR still above the scale mid-point of five) were for use of technology. It is noteworthy that in contrast to patterns for other items, IQRs suggested that participants did not endorse use of technology as highly on the ten-point scale in the final questionnaire. We explore this further in the qualitative analysis below.

Table 1. Participants' evaluations of the training materials, activities, structure, and technology (0-10 scale)

	Median	**IQR**	**Median**	**IQR**
Training materials	FQ1 (n=19)		Final FQ (n=19)	
Usefulness	7	7-8	8	8-8
Interest	7	7-7	8	7-8
User-friendliness	7	5-8	7	6.5-8
Quality	8	7-9	8	7-8
Training activities	FQ2 (n=22)		Final FQ (n=19)	
Usefulness	8	7-9	8	8-8
Interest	7.5	7-8	8	8-8
User-friendliness	7	6-8	7	7-8
Course structure	FQ3 (n=21)		Final FQ (n=19)	
Well-structured?	8	7-10	8	7-9
Clear?	8	7-9	8	7-9
Use of technology	FQ3 (n=21)		Final FQ (n=19)	
Usefulness	7	6-8	7	6-7.5
Supportiveness	7	6-8	7	5.5-7.5
User-friendliness	7	6-8	6	6-7

Participants' qualitative feedback on the item-writing course

We now present findings from the open-ended FQ questions (which provided explanations of scale ratings) and from interviews with trainees after the item writing course. Below, interview quotes are cited with participant number – P1, P2, etc; anonymous questionnaire responses are referred to as AQR. Eight themes were identified, discussed with data extracts below.

Item-writing theory

Participants generally reported that input on theoretical foundations of item writing was beneficial, particularly theory regarding test constructs and the features of different item types. In a characteristic statement drawn from the final interview, one participant said:

> "…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 [of test items]" (P16).

Some trainees also noted that they had not previously considered the difficulty and complexity of language at different levels – "before there was no real sense of grading

language" (P4); the course input on the CEFR appeared to help participants better understand what learners can do at different proficiency levels and, consequently, target the items they produced at the right level.

However, although participants generally found item-writing theory useful, the majority reported that they wanted theory to be presented in a concise and accessible way. PowerPoint presentations were the preferred input mode according to anonymous open-ended questionnaire responses, characterised as "effective", "clear", "applicable", and "interesting" (AQR). By comparison, the use of academic articles to generate discussion received mixed evaluations: while several participants found these informative and useful – "great introduction to important issues" (AQR) – most thought they were vague and heavy-going. Overall, only three trainees found input in the form of journal articles helpful. However, one participant suggested in the interview that optional readings could cater to those who wished to delve deeper: "there's one or two keen readers there, keenly attentive people that might be interested" (P23).

Example items

Participants particularly appreciated being provided with multiple item examples. For instance, a worksheet with 10 weak grammar MCQs introduced for analysis in Module 2 was described as "stimulating", "challenging", "very targeted", and "helpful" (AQR). One participant commented in the questionnaire that "this is exactly what I want from the course" (AQR) and requested more examples of both good and weak items – a sentiment repeated by most participants, in questionnaires and interviews:

> "I would've liked to have … more examples of good items versus poor item writing and the reasons behind them" (AQR)

> "… if I was suggesting improvements, I would say you can never have enough examples, even bad examples sitting next to good examples" (P24)

It also seems that one of the reasons participants valued the PowerPoint presentations was because they contained example items.

Item-writing practice

Practical item-writing activities were unanimously considered the most beneficial feature of the course. For instance, participants reported that writing a multiple-matching vocabulary task and producing speaking and writing prompts were the most useful activities in Modules 3 and 4 respectively. Participants believed that these activities were "challenging", "rewarding", and "interesting" (AQR), and generated useful feedback for future improvement. In the interviews, participants elaborated on the importance of item-writing practice: they believed that the theoretical principles and item-writing rules "become self-evident" (P2) while writing items. For example, P23 explained that "just reading about it [item-writing]" is insufficient because "you need to actually do it and get feedback about what I'm doing wrong".

Tutor- and peer-feedback

Individual tutor-feedback on items was perceived as crucial by most participants, who particularly appreciated feedback that was detailed, comprehensive, and followed a standard format. Tutors gave feedback within one week of each item-writing event and before a new round of item-writing practice. P11 found this continuous approach particularly helpful compared to online courses he had done before where "they expect everyone just to do their own thing and you get feedback way at the end". According to participants, an important benefit of this continuous feedback approach was that it allowed trainees to assimilate information before they attempted writing items again. For example, P9 said: "…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".

Attitudes to peer-feedback, however, varied 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" (AQR)

> "…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" (P8)

Others believed that peer-feedback had limited value because the discussion groups where participants reviewed each other's items varied in their levels of activity. In the interviews, participants elaborated on two problems with peer-feedback. First, the quality of feedback greatly depended on the individuals within the group: one pattern we observed was that the more active and forthcoming the participants, the more useful the peer-feedback was perceived to be. Second, because the groups changed in every module, many felt reluctant to give negative feedback to unfamiliar participants – as P24 put it, they "tended to pull punches".

Modes of interaction

Some activities of the item-writing training course required individual work, while others were done in groups. Quantitative data indicated that the combination of individual study and groupwork was the most-preferred mode of interaction (Table 4). The qualitative data confirmed this view; participants felt that it provided variety and reflected the different phases of real-world item writing, which combines solo, creative work with collaborative, review work:

> "[the combination of individual and groupwork] 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" (AQR)

Table 4. Participants' preferences for the mode of interaction (n=21)

	Working only individually	Working only in groups	A combination of both
Most preferred	7	1	13
2nd preferred	7	7	7
Least preferred	7	13	1

Besides peer-feedback on items, groupwork was reported as particularly suitable for activities involving analysis of example items: in most course modules, participants were given weak items to discuss as a group and to provide suggestions for improvement. Groupwork was, however, considered less suitable for learning about theoretical foundations of item writing. For example, P9 explained that the discussions worked better for activities that required "original and unique" contributions, whereas posting thoughts on an issue connected to item-writing theory was not as beneficial because once "one person wrote a lot of things at the beginning so there wasn't that much else to say".

Participants' preferences for group or individual work were, to some extent, dependent on beliefs about their own personal disposition: in questionnaire comments, some participants wrote that they generally preferred working alone while others felt they needed group collaboration to stay motivated. Some participants felt their groups were not active enough and suggested that groups should be moderated more closely to stimulate activity: a group leader should be appointed, or the tutors should be more active in encouraging individual participants to contribute.

Course structure

Among the four areas for feedback, course structure received consistently high endorsement from participants (Table 3). In their open-ended questionnaire comments, participants highlighted the following aspects of the course structure as particularly beneficial: (1) following a logical progression from theory to practice, and from simple to complex; (2) building on skills acquired in earlier modules, "so knowledge, conventions and skills acquired could be re-used" (AQR); (3) offering a balanced combination of "self-study, group discussions, sharing written items and peer reviews" (AQR); and (4) structuring each module in a similar way: "starting off from lead in, a form of narrated ppt, followed by independent then group tasks" (AQR). Participants valued a predictable module structure whereby they "generally knew what [they] were doing each week and why" (AQR).

Use of technology

Participants' comments helped clarify the reasons for the lower evaluations of the course technology compared to other aspects of the course (Table 3). Participants described the LMS Edmodo as "unsophisticated", "underwhelming", and "child-focussed" (AQR). Opinions on WeChat (used for group discussions and item peer-review) were almost equally split. Half of the respondents praised WeChat as a good platform for group activities, noting that it was appropriate in the Chinese context and worked very well. The other half, however, felt that WeChat was less suitable for group discussions due to WeChat's functionality specifics. Overall, WeChat was considered the most suitable for activities involving peer-feedback on items – "Wechat shines for this purpose" (AQR) – because items can be posted in discussion threads and all group members could discuss and comment on the items.

The use of email, as a familiar medium, did not pose any problem to participants and was thought to be "normal" and "appropriate" (AQR). However, several participants suggested eliminating email communication to reduce the range of technologies used on the course. They would have preferred all participant-tutor communication to happen within the LMS.

Suggestions for course improvement

When asked what other course materials they would have found helpful, most participants again suggested more example items, both good quality and faulty. Participants would have particularly preferred detailed explanations for why the items were considered high- or low-quality: "worksheets identifying errors in items with clear answers and explanations" (AQR). Additionally, participants asked for item-writing guidelines that detail the step-by-step item-writing process for items of each type and proficiency level. Another suggestion was presentations in which tutors would "talk through their mental process of creating an item" (AQR).

Although the course was already quite long (three months), five participants wanted the course to run longer with more fluid deadlines and more breaks between the modules to allow for catch-up. A longer course would also allow for "a second submission after the first QR review [item quality review] to really deepen the learning and have more feedback" (AQR).

Participants also provided suggestions on how technology might support their work during the course, for example, using Google Docs to write items as a group: "a co-written task (reading or listening) could be fun to put together" (AQR). While the course was taught fully asynchronously, several participants would have appreciated opportunities for synchronous communication: optional online webinars, online live Q&A sessions with course tutors, as well as using video-conferencing technology "to facilitate some pair-work item-writing" (AQR). One respondent additionally suggested including videos recorded by course tutors: "This video introduces a particular module and shares the course leaders' personal experience of being an item writer" (AQR).

Discussion

Quantitative results from the course feedback questionnaires indicated that trainees were generally satisfied with how the course was organised and delivered. The course structure was valued, and a positive appraisal of course materials and activities was maintained (and appeared to converge) as the course progressed. The delivery technology, however, was evaluated somewhat lower compared to other course aspects. Contextual factors may have influenced this: course participants were based in China while the lead tutor was in the UK and restrictions on foreign internet traffic and bans on some widely-used online platforms and apps affected the speed and quality of course access. Moreover, it should be noted that the course was delivered just before the Covid-19 pandemic, when online learning was less widespread, with few participants having experienced it before. Potentially, trainees' digital literacy – as reflected in their ability to access materials, follow online tutorials, and collaborate with other participants virtually – might have influenced their satisfaction with technological aspects of the training. However, clear lessons can be learnt from the feedback: online item-writing training courses should make use of LMSs that allow for collaborative groupwork and, ideally, for all trainee-tutor and trainee-trainee communication to happen within one system.

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. This provides support for Fulcher's (2012) principled position that theoretical knowledge forms an important foundation for developing practical skills and should be a regular part of item-writing training. As to how to incorporate the theory in training, Fulcher (2012) proposed doing this within the context of practical test construction, introducing theory along the way of test development. While this approach was not fully adopted in this training course (as it did not involve an actual test development cycle), theory was directly connected with concrete item writing activities and thus introduced in a contextualised manner. Input written in an academic style (such as journal articles) was generally perceived as less engaging, however, which suggests that theory should be introduced in more

accessible forms for this audience, such as brief presentations rather than academic publications. At the same time, several participants reported enjoying the academic readings and asked for additional literature. This indicates that item-writing training should aim to serve diverse types of trainees. One way of achieving this could be by including optional readings and tasks. These optional materials could include short, tutor-recorded video explainers or narrated slides for those who prefer visual or auditory input, alongside handbook-style, introductory-overview readings which are typically accessible to a wider audience (e.g., chapters in volumes of the Cambridge Language Assessment Series, in Fulcher & Harding (2022), Hughes & Hughes (2020), or Winke & Brunfaut (2021) and other relevant volumes in the Routledge Handbook Series on Second Language Acquisition). This can then be complemented with a curated list of research articles for trainees seeking deeper engagement with theoretical underpinnings. To maximise relevance, optional readings and tasks should be clearly linked to the session content but could also include exploratory materials for those interested in extending their understanding beyond the course's scope.

Participants unanimously wanted to see more items of each type, both good and problematic examples. The importance of multiple item examples for item writers has been advocated in the literature. For example, Popham (1994) suggested that item writers should be provided with "a set of varied, but not exhaustive, illustrative items" (pp.17-18), while Kim et al. (2010) wrote that "item writers need … a range of sample items with different difficulty levels" (p.165). However, provision of extensive sample items risks that item-writer trainees become over-reliant on examples at the expense of gaining a deep understanding of the principles underlying the production of valid test items. To mitigate this risk, training programmes should include a limited but varied set of sample items, selected to illustrate both successful features and a range of possible problems. Crucially, the rationale for including each sample item should be made explicit, linking it to specific item-writing principles so that trainees focus on the underlying constructs rather than simply replicating surface features. They should also be explicitly discouraged from overly relying on sample items.

Participants' feedback indicated that the practical nature of the training – regular item production and tutor feedback – was seen as the most useful course feature. This finding aligns with previous research on language assessment training more generally

(e.g., Harsch et al., 2021; Kremmel et al., 2018): trainees strongly prefer practice-orientated training over purely theory-orientated courses. In the field of educational measurement, item-writing training schedules proposed by both Downing (2006) and Welch (2006) also include item-writing practice and group discussions of items. However, the training length suggested in the literature thus far might not allow trainees to fully benefit from these activities. Haladyna and Rodriguez (2013), for example, believed that item-writing training should last from several hours to several days. Some trainees in the present study, however, preferred the course – already three months long – to last even longer. The trainees wanted opportunities to revise their items and receive a second round of feedback – something they felt would deepen their learning. This finding resonates with Koh's (2011) study, which compared the effectiveness of LAL training delivered as one-shot workshops and as long-term professional development, observing that the sustained long-term training was more effective. Long-term training is also more typical of item writer training and support in large-scale testing programmes, where professional item writers are often mentored, and receive feedback, over many months or years. Such extensive training now seems more feasible in online mode, which might suggest that online item-writing training would be preferable to face-to-face provision on the grounds of its temporal and geographical flexibility (or a blended format could be considered). Having several rounds of feedback-and-revision, however, as requested by some course participants, might be impractical even for online training. One solution might be to complement induction item-writing training with subsequent mentoring, where a newly-trained item writer is paired up with a more experienced one for feedback and ongoing support.

Trainees in this study valued prompt, continuous tutor feedback following each item-writing event. The timeliness and regularity of this feedback appeared to allow trainees to "absorb" feedback suggestions and apply them during item-writing practice for the following module. The course's structure seemed conducive to this: each new module built on the skills acquired in previous modules. Participants also appreciated feedback that followed a standard format (in this case, an item review checklist) and included detailed explanations for item evaluations. Overall, it seems that the above feedback characteristics – timeliness, thoroughness and a standardised format – are

viewed particularly positively by trainees. They also reflect key features of effective feedback as found in the general education literature (e.g., Winstone & Carless, 2020).

Although many trainees found peer-feedback useful, two problems were reported: not all peer-review groups were equally active, and some participants felt uncomfortable about giving negative feedback on others' items. Existing literature indicates that, despite its many advantages, peer-feedback can pose challenges in terms of reliability, especially when subjective judgements have to be made (O'Donnell, 1998), as is the case with item review. Another concern is social dynamics within the group which "might influence the reliability and validity of the assessments" (O'Donnell, 1998, p.263). Despite these drawbacks, the balance of evidence suggests that peer-feedback is a useful and necessary component of item-writing training, which can encourage a more reflective approach to item production, foster item-writer collaboration, and provide an extra layer of feedback on items beyond that given by course tutors. Ideally, to maximise peer-feedback usefulness, all trainees should feel part of an online learning community where they are responsible for each other's learning and where they feel safe to exchange honest feedback on items. Bos-Ciussi et al. (2008) formulated recommendations for cultivating a community of practice in virtual learning environments. First, tutors should stay in the background while at the same time "set up strict rules in order to encourage exchanges to emerge" (Bos-Ciussi et al., 2008, p.303). Moreover, learning content should be designed in such a way as to encourage students to interact. Interestingly, participants' suggestions in this study reflect the above recommendations: trainees wanted stricter group participation rules and possibly a group leader who would moderate group activities. Trainees also wanted fewer group rotations to allow more time for bonding.

The course structure received high evaluations from participants. The course was considered well-designed, well-paced, flexible, and clearly structured. Therefore, the following features, which characterised the course in this study, can be recommended for any 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 in a way that allows information chunks to build on each other and to be re-used later in the training.

This study's participants provided some suggestions for further strengthening item-writing training. Although the course already encouraged collaboration among trainees, participants wanted such collaboration to go further: several trainees requested item-writing sessions in pairs/groups using Google Docs or an online meeting facility. The importance of collaboration in item writing has also been highlighted in the literature. For example, Cui et al. (2022) found that collaborative test design practice by three EAP teachers resulted in better-quality tests and improved teachers' LAL overall. More generally, a combination of individual and collaborative work appeared to be valued because it mimics what might occur in real-world item writing. These findings also connect with literature (e.g., Green & Hawkey, 2011; Shin, 2022) which has noted the dual nature of item writing as an art and a science, involving both creative, solo work, and more collaborative, consensus building.

Conclusion

To the best of our knowledge, this study is the first published detailed empirical investigation of trainee perceptions of an item-writing training course. The findings support the generally-held assumption of the importance of item-writing training (Downing, 2006; Haladyna & Rodriguez, 2013; Welch, 2006). While this study was contextualised within an item-writing training course for experienced EFL teachers who were British Council employees (and novice item writers), the insights gained from participants' evaluations and feedback may be applicable to training item writers in a variety of contexts where practitioners may be looking for evidence-based recommendations on best-practice in item-writing training (see Pill et al., 2024). The study identified the following as likely to be particularly useful:

a) providing more general theoretical input and combining it with plenty of item writing practice;
b) prompt, comprehensive and detailed feedback on item writing practice;
c) multiple item review cycles, e.g. including peer-feedback rounds;
d) carefully managed/modelled group interactions to ensure the possibility of frank peer-feedback in a potentially face-threatening environment;
e) activities that require peer collaboration;

f) principled course design – cumulative progression, moving from simple to complex, and following-up input with discussion and practice; and
g) for online delivery, a unified training platform with functionalities for groupwork and peer reviewing.

We also remind the reader of the freely available course syllabus, specifications, and questionnaires:
https://osf.io/jsg3e/?view_only=73b49730049440 51a65bdb6b13cfca96

Further research into the usefulness of item-writing training is nevertheless recommended across different contexts to confirm or challenge the principles drawn from this study. This seems important given the highly contextualised nature of the study, and the fact that participants in other contexts may differ in their motivation to engage in training, their available time, and employer support, for example. Moreover, research into item-writing training would ideally go beyond investigations of trainees' perceptions. Quantitative measures of training effectiveness could look at changes in trainees' item quality before and after training. Indeed, as mentioned earlier, the present study was part of a larger project that also included a quasi-experimental pretest-posttest study whereby item quality before and after the training was evaluated by expert judges against an evaluation scale (Rossi, 2021; Rossi & Brunfaut, 2021). The field would benefit from more research focusing on the impact of item-writing training on measures of item quality.

Another interesting avenue for further research could involve an experimental design in which different item-writing training approaches/courses are directly compared (whereas the present study only looked into one, on its own). This might reveal variation in the effectiveness of different approaches in terms of item-quality improvements, trainee engagement, and/or course efficiencies such as time and resource requirements.

Finally, while this study did not involve using Generative AI (GenAI) for item writing, which was not prominent at the time of data collection, the rapid development of GenAI technologies is beginning to reshape the landscape of language assessment in general and item writing in particular. It would be sensible for future training programmes to cover components that help trainees to critically evaluate GenAI

capabilities and to ethically use GenAI tools - for example, for idea generation or drafting test items - while maintaining rigorous validity standards (Rossi, 2024; Rossi & Montcada, 2025). The first author's resource hub https://itemwriting.co/ includes a curated list of resources and publications on using GenAI for item writing.

Acknowledgements

We would like to thank the item-writing trainees and item reviewers who took part in our study, as well as the assistant tutor on the item-writing training course. We also thank the anonymous reviewers of this article and the SiLA editorial team for their supportive feedback.

Author disclosures

The authors acknowledge the role of the British Council in making this study possible. The British Council provided a research grant which enabled Rossi to conduct part of the study under the Assessment Research Awards and Grants programme 2018. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the British Council, its related bodies or its partners.

The authors also wish to declare the following recent and current British Council roles: Rossi has been an item editor and a consultant on several assessment projects. Brunfaut has been co-editor of the British Council Monographs on Modern Language Testing since 2021, and was a recipient of a 2022 Assessment Research Grant. Harding received grant funding from the British Council for the British Council-Lancaster Aptis Corpus project. He is also Co-Investigator on another British Council-funded project: The EMI Corpus project. He is currently a member of the British Council Assessment Advisory Board and has performed ad hoc consultancy for the British Council for over five years.

The authors' CRediT roles were: Rossi – Conceptualisation, Methodology, Data Collection, Data Analysis, Validation, Visualization, Writing-original draft, Writing-review & editing. Brunfaut – Conceptualisation, Methodology, Supervision, Writing-

original draft, Writing-review & editing. Harding – Data Analysis, Visualization, Writing-original draft, Writing-review & editing.

ORCID iDs

Olena Rossi https://orcid.org/0000-0002-6030-8168

Tineke Brunfaut https://orcid.org/0000-0001-8018-8004

Luke Harding https://orcid.org/0000-0001-9579-6571

References

Alderson, J. C. (2010). A survey of aviation English tests. *Language Testing, 27*(1), 51-72. https://doi.org/10.1177/0265532209347196

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.

Bachman, L. F. & Palmer, A. (2010). *Language assessment in practice*. Oxford University Press.

Baker, B. A., & Riches, C. (2018). The development of EFL examinations in Haiti: Collaboration and language assessment literacy development. *Language Testing*, *35*(4), 557–581. https://doi.org/10.1177/0265532217716732

Borg, S. (2021). Systemic in-service language teacher education. In E. Macaro and R. Woore (Eds.), *Debates in second language education* (pp.144-167). Routledge. https://doi.org/10.4324/9781003008361-10

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). Information Age Publishing.

Braun, V., & Clarke, V. (2012). Thematic analysis. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of*

research methods in psychology, Vol. 2. (pp. 57–71). American Psychological Association. https://doi.org/10.1037/13620-004

Brown, H. D., & Abeywickrama, P. (2018). *Language assessment principles and classroom practices* (3rd ed.). Pearson.

Brunfaut, T., & Harding, L. (2018). Teachers setting the assessment (literacy) agenda: a case study of a teacher-led national test development project in Luxembourg. In D. Xerri, & P. Vella Briffa (Eds.), *Teacher involvement in high stakes language testing* (pp. 155-172). Springer.

Cui, Y., Liu, Y., Yu, H., & Gao, Y. (2022). Developing English teachers' language assessment literacy in an EAP reform context through test design: A case study. *System*, *109*, 102866. https://doi.org/10.1016/j.system.2022.102866

de Jong, J. (2008, August). Procedures for training item writers and human raters. Paper presented at the *EALTA Annual Conference*, Athens, Greece.

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. https://doi.org/10.4324/9780203874776-6

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly, 9*(2), 113-132. https://doi.org/10.1080/15434303.2011.642041

Fulcher, G., & Harding, L. (Eds.). (2022). *The Routledge handbook of language testing* (2nd ed.). Routledge.

Giraldo, F., & Murcia, D. (2019). Language assessment literacy and professional development of pre-service language teachers. *Colombian Applied Linguistics Journal, 21*(2). https://doi.org/10.14483/22487085.14514

Green, A. (2016). Assessment literacy for language teachers. In D. Tsagari (Ed.), *Classroom-based assessment in L2 contexts* (pp.8-29). Cambridge Scholars Publishing.

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), 109–29. https://doi.org/10.1177/0265532211413445

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items.* Routledge. https://doi.org/10.4324/9780203850381

Harding, L., & Brunfaut, T. (2020). Trajectories of language assessment literacy in a teacher-researcher partnership: Locating elements of praxis through narrative inquiry. In M. Poehner & O. Inbar-Lourie (Eds.), *Towards a reconceptualization of second language classroom assessment: Praxis and researcher-teacher partnership* (pp. 61-81). Springer. https://doi.org/10.1007/978-3-030-35081-9_4

Harsch, C., Seyferth, S., & Villa Larenas, S. (2021). Evaluating a collaborative and responsive project to develop language assessment literacy. *Language Learning in Higher Education*, *11*(2), 311–342. https://doi.org/10.1515/cercles-2021-2020

Ho, E. C., & Yan, X. (2021). Using community of practice to characterize collaborative essay prompt writing and its role in developing language assessment literacy for pre-service language teachers. *System*, *101*, 102569. https://doi.org/10.1016/j.system.2021.102569

Huang, J. Y., Chan, H-Y., Lee, P-I., Tang, Y-W., Chiou, T-W., Chen Liu, K.C.S., & Ho, Y-F. (2022). Exploration of changes in pharmacy students' perceptions of and attitudes and attitudes towards professionalism: Outcome of a community pharmacy experiential learning programme in Taiwan. *BMC Medical Education, 22*, Article 195. https://doi.org/10.1186/s12909-022-03261-6

Hughes A., & Hughes, J. (2020). *Testing for language teachers* (3rd ed.). Cambridge University Press. https://doi.org/10.1017/9781009024723

Ingham, K. (2008). The Cambridge ESOL approach to item writer training: The case of ICFE listening. *Research Notes, 32*, 5-9. https://www.cambridgeenglish.org/english-research-group/published-research/research-notes/

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. https://doi.org/10.1097/00001888-200202000-00016

Karthikeyan, S., O'Connor, E., & Hu. W. (2019). Barriers and facilitators to writing quality items for medical school assessments – A scoping review. *BMC*

Medical Education, 19, Article 123. https://doi.org/10.1186/s12909-019-1544-8

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. https://doi.org/10.1080/15434300903473989

Knoch, U., Fairbairn, J., & Huisman, A. (2016). An evaluation of an online rater training program for the speaking and writing sub-tests of the Aptis test. *Papers in Language Testing and Assessment, 5*(1), 90-106. https://doi.org/10.58379/xdyp1068

Koh, K. H. (2011). Improving teachers' assessment literacy through professional development. *Teaching Education, 22*(3), 255–276. https://doi.org/10.1080/10476210.2011.593164

Kremmel, B., Eberharter, K., Holzknecht, F., & Konrad, E. (2018). Fostering language assessment literacy through teacher involvement in high-stakes test development. In D. Xerri & P. Vella Briffa (Eds.), *Teacher involvement in high-stakes language testing* (pp. 173-194). Springer. https://doi.org/10.1007/978-3-319-77177-9_10

Kvale, S., & Brinkmann, S. (2009). *InterViews: Learning the craft of qualitative research interviewing* (2nd ed.). SAGE.

Lam, R. (2015). Language assessment training in Hong Kong: Implications for language assessment literacy. *Language Testing*, *32*(2), 169–197. https://doi.org/10.1177/0265532214554321

Lowell, B. R., & McNeill, K. L. (2023). Changes in teachers' beliefs: A longitudinal study of science teachers engaging in curriculum-based professional development. *Journal of Research in Science Teaching, 60*(7), 1457–1487. https://doi.org/10.1002/tea.21839

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. https://doi.org/10.4324/9781315042527

Newby, A. C. (1992). *Training evaluation handbook*. Gower.

O'Donnell, A. M. (1998). Peers assessing peers: Possibilities and problems. In K J. Topping and S. W. Ehly (Eds.), *Peer-assisted learning* (pp.255-275). Lawrence Erlbaum.

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.

Osterlindt, S. J. (1998). *Constructing test items: Multiple-choice, constructed-response, performance, and other formats.* Kluwer.

Phillips, J. J. (1991). *Handbook of training evaluation and measurement methods.* Gulf Publishing. https://doi.org/10.4324/9780080572659

Pill, J., Bandman, J., Bottini, R., Brunfaut, T., Davidson, N., Davila Perez, G., Harding, L., Jung, Y., Lestari, S. B., Ramos Galvez, C., & Rossi, O. (2024). Shaping a language testing curriculum: Insights from an oral history of a Masters programme. In B. Baker & L. Taylor (Eds.), *Learning about assessing language: A matter of literacy or competency development* (pp.168-191). Cambridge University Press.

Popham, W. J. (1994). The instructional consequence of criterion-referenced clarity. *Educational Measurement: Issues and Practice, 13*(4), 15-18. https://doi.org/10.1111/j.1745-3992.1994.tb00565.x

Rae, L. (1999). *Using evaluation in training and development*. Kogan Page.

Rossi, O. (2021). *Item writing skills and their development: Insights from an online induction item-writing training course* [Doctoral thesis]. Lancaster University. https://doi.org/10.17635/lancaster/thesis/1267

Rossi, O. (2024, June). *Item writing with generative AI: Current issues and future directions* [Conference presentation]. EALTA Annual Conference, Belfast, UK.

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. https://doi.org/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, 18(4), 398-418. https://doi.org/10.1080/15434303.2021.1895162

Rossi, O., & Montcada, J. M. (2025, March). *Using ChatGPT to generate True/False reading comprehension items: Recommendations for practice* [Conference presentation]. Presentation EALTA's AI SIG Online Meeting.

Rubin, H. J., & Rubin, I. (2005). *Qualitative interviewing: The art of hearing data* (2nd ed.). SAGE. https://doi.org/10.4135/9781452226651

Shin, D. (2022). Item writing and item writers. In G. Fulcher & L. Harding (Eds.), *The Routledge handbook of language testing* (2nd ed.) (pp.341–356). Routledge. https://doi.org/10.4324/9781003220756-27

Spaan, M. (2007). Evolution of a test item. *Language Assessment Quarterly*, *4*(3), 279-293, https://doi.org/10.1080/15434300701462937

Steffe, L.P., & Gale, J. (1995). *Constructivism in education*. Lawrence Erlbaum. https://doi.org/10.4324/9780203052600

Tsagari, D., & Vogt, K. (2017). Assessment literacy of foreign language teachers around Europe: Research, challenges and future prospects. *Papers in Language Testing and Assessment*, *6*(1), 41-63. https://doi.org/10.58379/uhix9883

Villa Larenas, S., & Brunfaut, T. (2023). But who trains the language teacher educator who trains the language teacher? An empirical investigation of Chilean EFL teacher educators' language assessment literacy. *Language Testing*, *40*(3), 463-492. https://doi.org/10.1177/02655322221134218

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. https://doi.org/10.4324/9780203874776.ch13

Winke, P., & Brunfaut, T. (Eds.). (2021). *The Routledge handbook of second language acquisition and language testing*. Routledge. https://doi.org/10.4324/9781351034784

Winstone, N., & Carless, D. (2020). *Designing effective feedback processes in higher education: A learning-focused approach*. Routledge. https://doi.org/10.4324/9781351115940

Yan, X., & Fan, J. (2021). "Am I qualified to be a language tester?": Understanding the development of language assessment literacy across three stakeholder groups. *Language Testing, 38*(2), 219-246. https://doi.org/10.1177/0265532220929924