RESEARCH

Open Access

Investigating the characteristics of language test specifications and item writer guidelines, and their effect on item development: a mixed-method case study

Zahra Ali Al Lawati[1]*

*Correspondence: zahral@squ.edu.om

[1] Centre for Preparatory Studies, Sultan Qaboos University, Muscat, Oman

Abstract

This study discusses the characteristics of test specifications (specs) and item writer guidelines (IWGs), their role in item development of English as a Second Language (ESL) reading tests, and the use of the CEFR for specs development. This mixed-method study analyzed specs, IWGs, tests, and the Pearson Test of English General test statistics. Moreover, interviews and focus groups were conducted with the specs' developers, IWGs, and item writers. The findings show no unique way of conceptualizing specs and IWGs. Moreover, translating the CEFR reading descriptors into specs is a challenging task. However, results from the judgmental study and item statistics suggest that the investigated specs and IWGs facilitated the development of good-quality items at a certain difficulty level. This study reveals the potential role of specs and IWGs in establishing test validity. This research contributes to understanding the under-researched area of specs and IWGs and shows the type of information required for effective item writing and ways of enhancing the validity and reliability of tests. Practical and theoretical suggestions and future research have also been identified.

Keywords: AUA, CEFR, ESL reading tests, Item writer guidelines, Pearson Test of English General (PTE General), Test specifications

Introduction

Language testing researchers have emphasized the importance of using precise specifications (specs) for test development (e.g., Davidson & Lynch, 2002; Fulcher & Davidson, 2007; Fulcher et al., 2022; Fulcher, 2021a, 2021b; Green & Hawkey, 2011; Hughes, 1989; Jin (2021); Norris, Brown, Hudson & Yoshioka, 1998). One of the early supporters of developing detailed specs was Hughes (1989), who argued: “that the essential first step in testing is to make oneself clear about what it is one wants to know and for what purpose” (p. 48). Researchers seem to agree that test specs must be clear and explicit to assist item writers in the item-writing process. However, there is a lack of research on test specs, and the details item writers require to produce the intended items. The few studies found are Kennedy (2007), Belyazid (1996), Li (2006), and Cho (1995). All four studies expanded our insights into test specs by researching aspects of specs development and use.

Regarding item writer guidelines (IWGs), there do not seem to be definitions in the literature of this term. Additionally, the topic of IWGs has hardly any coverage in the language assessment literature. The literature does not describe the exact role of the IWGs document, the reasons for its development, how to use it, and how it differs from the specs document. Therefore, this study addresses this gap in the literature by exploring specs and IWGs, the details provided in these documents, and the relation between specs, IWGs, and produced items.

Literature review

Very little has been published on test specs, but "Test craft" by Davidson and Lynch (2002) forms an exception. In their book, entirely dedicated to specs, Davidson and Lynch (2002) define specifications broadly as "a generative blueprint from which test items can be produced" (p. 4). The specs document is thought to be essential in the test development process. It is believed to promote stakeholder discussions about what the test intends to assess. According to Davidson (2012b), operationalizing target skills and abilities designed to be assessed into measurable terms is a challenge, and the specs are the central place for doing this. Thus, specs are thought to translate the test developers' construct into content descriptions, which the item writers then translate into actual items.

Similarly, Bachman and Palmer (2010) discuss specs' vital role in an Assessment Use Argument (AUA) framework. At the assessment development level, specs could be used to support the AUA warrants, which show that test items match the intended items as stipulated by the developers at the design stage. At the assessment justification level, Bachman and Palmer (2010) assert that each specs' component addresses one or more warrants in the AUA. They argue that specs assist in evaluating the link between the test developer's intentions (described in the specs) and the tests produced.

Test specs include different components, which vary in detail depending on the exam purpose and the examination body. Several specs formats have been described in the field, such as Davidson and Lynch (2002), Alderson et al. (1995), Bachman and Palmer (1996), Norris (1998), and Tinkelman (1971). Most specs documents share many components, and differences often lie in the level of detail provided. This paper, however, considers only Davidson and Lynch's (2002) format.

Davidson and Lynch (2002) proposed a five-component spec format, which they adapted from the work of Popham (1978). The General Description (GD) provides information on what is intended to be assessed. It may cover many testing details, such as the skill that should be tested and the reason for measuring that skill; the Prompt Attributes (PA) provide instructions on what should be presented to the test-takers to tell them what they are supposed to do. The PA includes information on the item types to be used to test the target skills; the Response Attributes (RA) clearly express how the test-takers will need to respond to the items; the Sample Item (SI) exemplifies a translation of the descriptions given in the earlier sections of the specs into the intended items, and the fifth component is the Specification Supplement (SS). Sometimes extra information is provided to specs users to help them use the document in the intended way. Davidson and Lynch (2002) explain that the SS is not a compulsory section. However, it can be an excellent place to provide the specs users with as much detail as possible without overloading the other components. They thus suggest that a certain level of detail may be needed for specs users, but providing much information in the specs can cause the document to be complicated.

There are two main procedures for developing specs described in the literature. The first is where the specs are constructed to create a new test. The second method is called "Reverse Engineering" (Davidson & Lynch, 2002; Davidson, 2012a). Here, specs are drawn up or linked to developed tests after developing a test. There can still be differences in the starting point within these two main procedures, going from the overall test to items, from the items to the overall test, or everything together. However, research on which manner is most effective or valid is lacking. In addition, despite the importance researchers have attached to test specs (e.g., Alderson et at., 1995; Bachman & Palmer, 2010; Davidson & Lynch, 2002; Tinkelman, 1971), surprisingly, hardly any research has been conducted on specs and IWGs. Consequently, the literature on developing these documents, their use, and whether they may lead to better item writing is scarce. Although, according to Shin (2021), the actual process of how items are written from test specs is an area of active enquiry, very little is known about these two documents' characteristics and their relationship.

The only studies found that specifically focus on test specs are post-graduate research by Kennedy (2007), Belyazid (1996), Li (2006), and Cho (1995). All four studies expanded our insights into test specs by researching aspects of specs development and use. However, the first three studies did not illuminate the specs' characteristics and the type of detail needed in this document by item writers to facilitate the development of the intended items. Cho (1995) is the only study investigating the effect of the specificity of specs on item construction. However, it did not explore the processes that the item writers followed for producing items, the details they paid attention to in the specs, and whether the specs can ensure that items are produced at a particular difficulty level. Similarly, Gutiérrez Baffil & Collada Peña, (2022) investigated the process of developing rating scales for writing according to test specifications and item writer guidelines. Another study reports a development and validation project for assessing writing by developing and validating local writing checklists (Harsch & Seyferth, 2019). Rossi and Brunfaut (2021) explored the effectiveness of an existing item-writing training course to produce authentic-sounding listening texts within the constraints of test specifications. Arhin et al. (2021) focused on item writing flaws in a communication skills test. However, these studies are not concerned with the reading tests, which this study hopes to address.

Another reason for this study is that "there is little information about planning, designing, and writing test items" in the literature (Osterlind, 1998, p. 3). The lack of research on constructing test items has long been acknowledged. As far back as 1951, Ebel suggested an extreme dearth of research on item writing. The list of researchers who have expressed similar observations also includes Green and Hawkey (2011), Kim et al. (2010), Roid and Haladyna (1982), and Salisbury (2005).

Kim et al. (2010) study is the first to explore the process of using specs while developing items in detail. However, the study did not describe the process of developing the specs (versus using these) and the kind of information included in the document.

Similarly, Green and Hawkey's (2011) study is very valuable since it investigated the role of specs and item writers in item development. However, their study focuses on reading text development and does not discuss item development processes.

Recently, there has been some realization about the importance of training item writers. Shin (2021) discusses the role of item writers in ensuring fairness and validity in language testing and emphasizes the importance of providing training and support for item writers and establishing clear guidelines and standards for item writing. Bafill (2022) assigns a lack of resources and training for teachers producing low-quality writing assessments. Haladyna and Rodriguez (2021) propose using full-information item analysis (FIIA) training for item writers to improve the quality of items produced. The study conducted by Arhin, Essuman and Arhin (2021) found that many test items contained one or more item writing flaws (such as ambiguity, irrelevant difficulty, and lack of specificity). They discuss the implications of these findings for test design and suggest that teacher training in item writing may be necessary to improve the quality of test items. Rossi and Brunfaut (2021) investigate whether item writers can be trained to produce authentic-sounding texts for listening assessments. The study involved a training program for item writers and found that item writers should receive ongoing training to produce reliable and valid items. Jin (2021) emphasizes the importance of test specifications in careful planning and development while creating a language proficiency test. Jin (2021) proposes a four-step process for developing test specifications, which includes defining the test's purpose, identifying the target population and their language needs, designing the test format and tasks, and validating the test through piloting and statistical analysis. In spite of this development, where the focus is on item writer training and test specifications, the conceptualization of test specs and IWGs and what constitutes test specs and IWGs has remained under-researched.

The literature review on defining item writer guidelines reveals no definitions of this term in the literature, although we find some examples of IWGs. In other words, the topic of IWGs is not adequately covered in the literature. For example, although Alderson (2000) and Alderson et al. (1995) recommend that one of the stages of test development is the use of guidelines for the training of item writers, no details are provided on how to develop IWGs, what should be included in such a document, or how to use it. When dealing with IWGs, Alderson et al. (1995) present some of the most frequent problems related to developing and using objective and subjective item types. The same is the case with the other documents that deal with IWGs (e.g., Alderson & Cseresznyés, 2005; Haladyna et al., 2002; Hambleton & Eignor, 1979). These documents provide item writers with instructions and tips on constructing different item types and problems that should be avoided while developing particular test items. However, the literature does not describe the exact role of the IWGs document, the reasons for its development, how to use it, and how it differs from the specs document. Therefore, this study addresses this gap in the literature by exploring the IWGs document, its details, and the relationship between the IWGs, specs, and produced items.

The current study

As argued above, test specs are an essential document in the assessment process and in establishing test validity. However, little research has been conducted on specs and

IWGs. Thus, this study examines the two documents more closely through a case study. More specifically, it investigates how the documents are developed and operationalized, in what respects they differ from one another or are similar, and the kind of information needed in these documents to help item writers develop the intended items.

Based on the above discussion on the importance of specs and IWGs, and the gaps identified in the literature on these documents, the following research questions (RQ) were developed:

> RQ 1: What are the test specifications developers' views on the feasibility of translating the reading CEFR descriptors into reading test specifications?
> RQ 2: What is the difference between the reading test specifications and reading item writers' guidelines in the investigated organization as conceptualized by the item writers and the developers of the two documents?
> RQ 3: What information do item writers need in reading test specifications and reading item writer guidelines to help them develop the intended reading items?
> RQ 4: To what extent do items developed based on reading test specifications and reading item writer guidelines match the intended Common European Framework of Reference (CEFR) levels and the intended statistical performance features?

A case study approach has been adopted to examine the above research questions. Within this approach, mixed-methods research (Stake, 2008; Tashakkori & Teddlie, 2003) was chosen, whereby data collection and analysis constitute a combination of qualitative and quantitative methods. The investigated specs and IWGs in this study are developed and used by Pearson Language Tests for constructing the Pearson Test of English General (PTE General). The PTE General is a suite for assessing and certifying the general English language ability of ESOL learners. Previously known as the London Tests of English, the revised PTE General has six levels (Level A1 to Level 5) linked to the six levels of the CEFR. The PTE General is intended for 14 years and older learners, assesses communication skills, and confirms English language learners' progress in learning English (Pearson, 2012). This study focuses on the reading part of the test, which consists of four sections: gap-fill multiple-choice, graphical multiple-choice, open-ended short answer questions, and note/text completion items. The investigated PTE General specs and IWGs are confidential documents that this paper cannot reproduce.

Methods used in this case study

Previous language testing research focused on test specs (Belyazid, 1996; Cho, 1995; Kennedy, 2007; Li, 2006) and used one or two data collection and analysis methods. However, this study adopted a mixed-method approach to enhance the study's construct validity. Therefore, a combination of qualitative and quantitative approaches for data collection and analysis was triangulated to obtain insights from different sources and groups of participants. Thus, this case study includes the following research instruments: a focus group with item writers, individual interviews with the developers of specs and IWGs, individual interviews with item writers, two steps of the judgmental study, and test statistics. Figure 1 comprehensively describes this study's qualitative and quantitative methods.

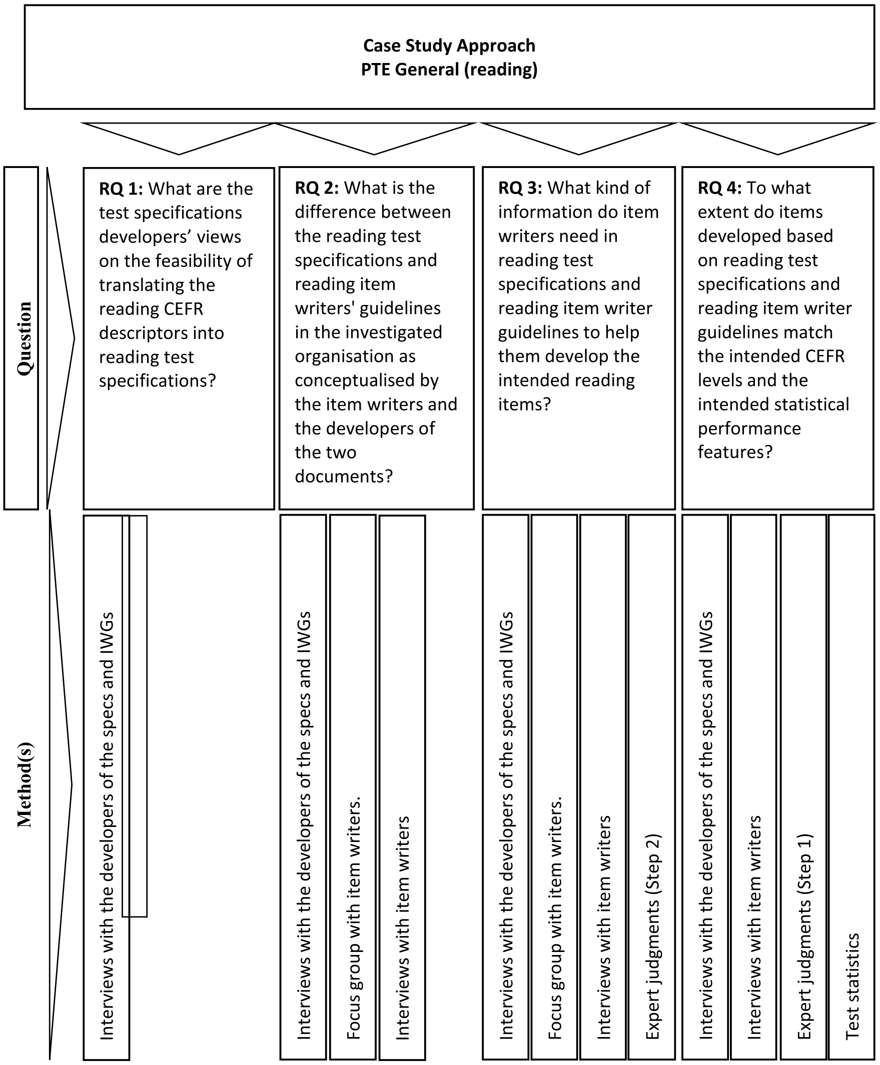
Case Study Approach
PTE General (reading)
Question
RQ 1: What are the test specifications developers' views on the feasibility of translating the reading CEFR descriptors into reading test specifications?
RQ 2: What is the difference between the reading test specifications and reading item writers' guidelines in the investigated organisation as conceptualised by the item writers and the developers of the two documents?
RQ 3: What kind of information do item writers need in reading test specifications and reading item writer guidelines to help them develop the intended reading items?
RQ 4: To what extent do items developed based on reading test specifications and reading item writer guidelines match the intended CEFR levels and the intended statistical performance features?
Method(s)
Interviews with the developers of the specs and IWGs
Interviews with the developers of the specs and IWGs
Focus group with item writers.
Interviews with item writers
Interviews with the developers of the specs and IWGs
Focus group with item writers.
Interviews with item writers
Expert judgments (Step 2)
Interviews with the developers of the specs and IWGs
Interviews with item writers
Expert judgments (Step 1)
Test statistics

Fig. 1 Overview of the methods used in this study

Qualitative methods

Several qualitative methods were employed in this study. To begin, a qualitative analysis of the PTE General reading specs, IWGs, and reading tests was conducted. The documents were examined at face value in terms of content and according to the following criteria: clarity, coherence, organization, and user-friendliness. This analysis aimed to gain initial insights into the nature and clarity of these documents and the extent to which they potentially facilitate the production of the intended PTE General reading test items. In addition, this preliminary analysis informed the development of the different data collection instruments used in this study. A second qualitative method concerned a focus group with seven PTE General item writers. It was used to gain insights into the clarity of the PTE General specs and IWGs these item writers use to develop PTE General reading test items and the processes they employ when developing them—findings from the focus group fed into RQs 2 and 3. In addition to the focus group, individual interviews were conducted with four item writers who had not participated in the focus group. This was done to understand better the process item writers employ and the documents they use when developing PTE General reading items. This was done to contribute to the answers to RQs 2, 3, and 4. The interview method was also employed to gather data on the process that the developers of the specs and IWGs went through while designing these two documents and on the difficulties they may have experienced in doing so. The data collected in these interviews contributes to all four RQs.

Quantitative methods

The above-described qualitative methods were complemented by quantitative data gathering and analyses. More specifically, to study the match between the PTE General specs and IWGs and the level of difficulty of items developed by PTE General item writers, test and item statistics provided by the exam board were analyzed in terms of item difficulty and item discrimination. This analysis helped to answer RQ 4.

A judgmental study was conducted by five language testing experts to examine the extent to which the PTE General specs and tests are linked to the CEFR. The judgmental data contributed to the answers to RQs 3 and 4.

Results

This section summarizes the findings obtained in this study and will be discussed in the next section. Firstly, in response to RQ1, the feasibility of using the reading CEFR descriptors for developing test specs was investigated through individual interviews with the document developers. They reported that linking the CEFR descriptors to the test objectives is challenging. The reasons they gave (the CEFR was not designed to describe testing, there are places in the CEFR descriptors where terminology is vague, there is some ambiguity in the CEFR descriptors, etc.) were comparable to those identified in other studies that used the CEFR for test development (Alderson et al., 2006; Davidson & Fulcher, 2007; Huhta et al., 2002; Jones, 2002; Weir, 2005).

For RQ2, views of the item writers and the document developers were analyzed to find the differences between specs and IWGs. All participants agreed that the IWGs is the item writers' main document and should be more detailed than the specs. The developers and the focus group item writers conceptualized the specs and the IWGs as separate documents designed for different purposes and users. However, the item writers in the individual interviews thought there was much overlap between the two documents and believed that the specs should be embedded in the IWGs. This stance of the participants does not support the literature on the purpose of the specs document. For example, Alderson et al. (1995) explain that specs are developed for different users and not just item writers. Moreover, the specs' content and format vary depending on the document's users.

RQ3 investigated the information needed in specs and IWGs to produce the intended items. Data from the following methods were analyzed: a focus group and interviews with item writers, interviews with the developers, and step 2 of the judgmental study. Findings related to RQ3 were classified under the following five headings:

Types of details

The item writers indicated they needed practical details to help them produce the intended items. By practical details, they meant specifications of text topics, text types, text length, objectives, time limits, candidates' age and background, item types, answer keys, grammatical syllabi, vocabulary lists, sample items, and guidelines on developing the different item types. The key pieces of information that the item writers reported referring to the most are sample items and word count. In addition, they reported that they would prefer to have one primary document, which is not very long.

Test objectives

Three developers perceived the objectives and the CEFR descriptors necessary for item writing. However, most item writers seemed to find the objectives more critical than the CEFR. They thought they had a good "feel" for the CEFR levels and that the objectives were more transparent and more accessible to use than the CEFR.

Sample items

When developing items, item writers believe they need sample items that are vital information. They reported that sample items are either the first piece of information they look at regarding the specs or the component they refer to the most in the process of item development. In addition, all item writers emphasized providing good sample items that carefully and comprehensively follow the requirements stipulated in the specs and IWGs.

Specific details

The item writers found it difficult to adhere to many details and requirements to produce the intended items. In addition, they indicated that developing, finding, and adapting texts seemed to be particularly difficult, especially for specific item types at certain levels. Findings from the following supported views on the difficulty of developing texts: the history of items and the specific requirements' parts (A and B) of step 2 of the judgmental study.

Training

All item writers considered training essential in item development and asked for more specific training. In addition, they made suggestions for future training sessions such as (1) using the specs and IWGs in training sessions to show item writers how to use these when developing items, (2) providing level-specific training, (3) more training on the CEFR levels and scales, and (4) training on finding suitable texts, text types, and texts suitable for the different item types.

For RQ4, step 1 of the judgmental study and item statistics analyses were conducted. Findings indicated that except for the highest test levels in many cases, the experts' judgments were proportionally either the same or differed by one level from the intended CEFR level.

To inform the answer to RQ4, item statistics of the 100 items included in the study were analyzed by examining the following:

- The level of difficulty by inspecting the *p*-values of items.
- The item discrimination through studying the point-biserial correlation of each item with the overall reading and test scores.
- The frequency with which keys and distractors were chosen in multiple-choice questions.

The above statistics were selected since these were shared by the exam board and standards used by the exam board for item analysis. The item difficulty and discrimination statistics for the 100 items were compared with the acceptable range statistics used by the exam board as indicators of item quality and difficulty level. Table 1 summarizes the number of items per test level and the number of candidates the data is based on.

Tables 2, 3, 4, 5, and 6 show that item statistics met the specified requirements.

Based on the above analyses, it can be concluded that, overall, the item's statistics aligned with the requirements specified by the exam board:

- About 64% of the items met the item difficulty range specified by the exam board.
- Around 88% of the item discrimination statistics (calculated for each item with the reading sub-test score) met the specified requirements.

Table 1 The 100 items included in the study

	Level 1	Level 2	Level 3	Level 5
Number of items	25	25	25	25
Number of candidates	406	3475	6207	1447

Table 2 The extent to which the 100 items meet the specified difficulty range (based on *p*-value and organized according to the test levels

	Met specified range	Above specified range	Below specified range	Total
Level A1	13 (52%)	11 (44%)	1 (4%)	25 (100%)
Level 2	18 (72%)	5 (20%)	2 (8%)	25 (100%)
Level 3	16 (64%)	4 (16%)	5 (20%)	25 (100%)
Level 5	17 (68%)	6 (24%)	2 (8%)	25 (100%)
Total	64 (64%)	26 (26%)	10 (10%)	100 (100%)

Table 3 The extent to which the 100 items meet the specified difficulty range (based on *p*-value and organized according to the test sections)

	Met specified range	Above specified range	Below specified range	Total
Section 4	15 (75%)	3 (15%)	2 (10%)	20 (100%)
Section 5	14 (70%)	5 (25%)	1 (5%)	20 (100%)
Section 6	17 (53%)	11 (34.4%)	4 (12.5%)	32 (100%)
Section 7	18 (64%)	7 (25%)	3 (11%)	28 (100%)
Total	64 (64%)	26 (26%)	10 (10%)	100 (100%)

Table 4 The extent to which the 100 items meet the specified discrimination index (based on point-biserial correlations) for the reading sub-tests per test level

	Met specified value	Below specified value	Total
Level A1	20 (80%)	5 (20%)	25 (100%)
Level 2	23 (92%)	2 (8%)	25 (100%)
Level 3	22 (88%)	3 (12%)	25 (100%)
Level 5	23 (92%)	2 (8%)	25 (100%)
Total	88 (88%)	12 (12%)	100 (100%)

Table 5 The extent to which the 100 items meet the specified discrimination index (based on point-biserial correlations) for the entire test per test section

	Met specified value	Below specified value	Total
Section 4	12 (60%)	8 (40%)	20 (100%)
Section 5	11 (55%)	9 (45%)	20 (100%)
Section 6	23 (72%)	9 (28%)	32 (100%)
Section 7	24 (86%)	4 (14%)	28 (100%)
Total	70 (70%)	30 (30%)	100 (100%)

Table 6 The extent to which the multiple-choice distractors and answer keys meet the specified item statistics criteria

	Key's frequency		Distractors' balance	
	Highest frequency	Not highest	Evenly balanced	Not balanced
Section 4	18 (90%)	2 (10%)	12 (60%)	8 (40%)
Section 5	20 (100%)	-	10 (50%)	10 (50%)
Total	38 (95%)	2 (5%)	22 (55%)	18 (45%)

- Approximately 70% of the item discriminations (calculated for each item with the entire test score) met the specified requirements.
- Finally, about 95% of the MC-key frequencies met the specified rules.

Although this analysis is based on 100 items, the number of candidates it is based on is significant, as shown in Table 1. Therefore, one can argue that the data reflects a real-life administration size and can give a clear picture of the quality of the items investigated. Thus, to a certain extent, the above findings may suggest that the reading specs and IWGs facilitated the development of items at a certain difficulty level and good quality in terms of discriminatory power.

Discussion

This section discusses key findings concerning previous research to explore the contributions of this study. The discussion is organized by research questions: the feasibility of translating the reading CEFR descriptors into reading test specs (RQ1), the difference between reading specs and reading IWGs is explored from the point of view of the document developers and the item writers (RQ2), investigating the kind of information item writers need in reading specs and reading IWGs to help them develop the intended reading items (RQ3), and the extent to which items developed based on reading specs and reading IWGs match the intended CEFR levels and the intended statistical performance features (RQ4).

Translating the reading CEFR descriptors into reading test specifications: feasibility

The specs developers concluded that using the CEFR to develop specs and linking the CEFR descriptors to the test objectives was not an easy task (Alderson et al. (2006), Davidson and Fulcher (2007) and Weir (2005). Davidson and Fulcher (2007) thought that the CEFR "lacks the necessary details on which to build test specifications" (p.232). Weir (2005) concluded that "though also containing much valuable information on language proficiency and advice for practitioners, in its present form, the CEFR is not sufficiently comprehensive, coherent or transparent for uncritical use in language testing" (p. 281). Therefore, the difficulties perceived by the developers of the specs in this study are not innate to this study but were experienced in other studies as well: using the CEFR for developing test specs is not straightforward.

The specs developers gave the reasons as to why they found using the CEFR for developing test specs a difficult task was also comparable to the reasons reported by others who had drawn up specs using the CEFR (Alderson et al., 2006; Huhta et al., 2002; N. Jones, 2002; Little, Simpson, & O'Connor, 2002; Morrow, 2004). For example, Alderson et al. (2006) identified four issues related to using the CEFR to build specs. These match well with the reasons given by the specs developers in this study.

Another limitation of the CEFR discussed by the document developers is the lack of description of the text and task features suitable for different descriptors at the different CEFR levels, which they had to infer. Information on the text and task characteristics at different levels (e.g., text length, what test-takers are intended to do with texts, reasons for reading a text, typical vocabulary, and structures) is either not defined or is ill-defined in the CEFR (Alderson et al., 2006; Weir, 2005).

To sum up, the document developers' responses concerning RQ1 confirmed the literature. They perceived using the CEFR descriptors to describe testing objectives and linking the CEFR descriptors to the test objectives to be challenging. The explanations given by the developers for their perceptions were comparable to the reasons reported in other studies that used the CEFR for developing tests. In addition, the item writers discussed similar difficulties with using the CEFR descriptors to develop reading test items. Thus, to enhance the feasibility of using the CEFR for testing, it needs to be made more comprehensive and transparent to enable testers to develop specs and tests that assess the intended constructs and, as a result, to be more accountable to stakeholders.

Reading test specifications and reading item writers guidelines: differences between the document developers and the item writers

RQ2 focuses on the differences between specs and IWGs from the point of view of the document developers and the item writers. Thus, it is discussed from two perspectives: practical, i.e., what the participants thought the differences between the two documents are, and theoretical, i.e., the role of specs in test validation. From the practical perspective, the participants had a general agreement about defining test specs and IWGs.

However, the focus group with item writers and interviews with the document developers indicated that the participants believed the specs and the IWGs served different purposes and users. The diverging opinions of the focus group and the interviewed item writers could be associated with the nature of the focus group method since one of the limitations of this method is that the data might depend on who is present in the group (Morgan, 1997). Thus, it could be that the focus group item writers were influenced by what the more dominant and experienced item writers said.

The above discussion thus contributes to the language testing field by suggesting that no fixed relationship exists between the specs, the IWGs, and their content. According to the testing literature, depending on the exam and the testing context, specs and IWGs could take different formats. It could also be a matter of terminology and labeling, but the necessary content is included in the document(s) in some way or another. In practice, in this case study, the specs and IWGs are separate documents, but there are individuals (some item writers) who thought the two documents could be merged into one and that there was no need for the specs. However, it has been shown that according to the developers and some item writers, the specs and IWGs can be conceptualized as two documents for different purposes and users: the specs for other users and the IWGs for item writers.

Furthermore, this study adds to the literature that, within the same assessment context, there are different views between individuals involved in test development: the developers versus the item writers and even among the item writers. Therefore, depending on the exam, the exam board, the assessment context, or the individuals involved, the specs and IWGs may be conceptualized as two separate documents, as two versions of the same document, or as one comprehensive document. This suggests that practitioners in the field may not need to worry about having one or two documents if the key information is included in the document(s).

Information needed in reading test specifications and reading item writer guidelines to help item writers develop the intended reading items.

RQ3 aimed to investigate the information needed in specs and IWGs to facilitate the production of the intended items. Data for RQ3 was collected using a focus group with item writers and individual interviews with developers and item writers. In addition, step 2 of the judgmental study was conducted to obtain experts' judgments on whether the items meet the requirements stipulated in the specs and IWGs.

The study shows divergent views of the document developers and item writers on the kind and level of detail needed, suggesting that it would be valuable to involve item writers in developing documents such as specs and IWGs. This may also help enhance the face validity of the documents for the item writers. Item writers are the ones who translate specs and IWGs into test items, and involving them from the early stages of developing such documents may be of benefit. Their views and ideas about details needed for developing the different item types may facilitate the production of the intended items and, consequently, enhance test validity. Involving item writers in developing such documents has also been suggested by Salisbury (2005) and Kim et al. (2010).

Based on the findings for RQ3, it can be argued that the CEFR descriptors should also be included in the IWGs for three reasons. Firstly, most developers considered both the objectives and CEFR descriptors important for item writing and for enhancing test validity by developing the intended items in terms of type and difficulty level. Secondly, they designed the IWGs as the item writers' principal document. Thus, it seems logical to expect that it contains all details the developers consider essential, which include the CEFR descriptors. Thirdly, by including the CEFR descriptors in the IWGs, it is possible that the item writers' attention will be directed more to their importance and, consequently, they may use them more when developing items. This may enhance their ability to develop the intended items and, as a result, strengthen the test's validity.

Therefore, test developers and testing organizations need to ensure that item writers know the importance of information on constructs and what items are supposed to assess to facilitate the production of the intended items. This could be done through the documents provided and training.

In sum, writing the intended items may not solely depend on providing item writers with well-developed documents. Other factors may come into play in the process of item development. One is how item writers use or interpret the documents provided. As discussed earlier, the item writers in this study employ different strategies and use the documents differently. Thus, training item writers on the intended use of the documents might facilitate their production of the intended items.

Match between items developed based on reading test specifications and item writer guidelines, the intended CEFR levels, and the intended statistical performance features

The key data for answering RQ4 is the analysis of item statistics and step 1 of the judgmental study. This was triangulated with data collected through individual interviews with the item writers and the document developers.

Although there were instances of disagreement in the group judgments, especially at the higher test levels, a still reasonable agreement was obtained in this study. Relating this amount of agreement to the judges employed and their experience in giving the type of judgments required suggests that experience and practice are essential in making judgments. Therefore, an interpretation is that the reading specs and IWGs investigated in this study tended to facilitate the development of items judged at or closer to the intended CEFR levels. However, variables other than the quality of the specs and IWGs (e.g., item writer skills and experience, training, feedback, rounds of editing and revision) may have contributed to the match between the items and the intended difficulty level.

The study shows a divergence between judgments on difficulty levels as opposed to empirical difficulty levels probably shows that item difficulty may not depend on item characteristics alone. Instead, item difficulty could result from text/item characteristics, test-taker characteristics, and the interaction between these (Bachman, 2002). Thus, empirical difficulty levels could result from interactions between item characteristics and test-taker characteristics. In contrast, judgments may be based on item characteristics and how the judges perceive test-takers would process the items. However, test-takers may interact with the items differently than expected and use different processes than those predicted by the judges. This does not mean that judges' judgments are not helpful. However, more research is needed on comparing the processes used by judges when making judgments with the process followed by test-takers while performing test tasks to improve judges' judgments potentially.

In general, item statistics met the requirements specified by the exam board. Indirectly, this suggests that the reading specs and IWGs facilitated or helped ensure that items at a certain difficulty level and of good quality in terms of discriminatory power were developed. However, it must be kept in mind that the items investigated have gone through rounds of revision and editing. Thus, these desired statistics could result from the quality of the documents, the skills and experience of the item writers, the quality of the revision process, or a combination of these.

The analysis of item statistics showed that overall, the item statistics met the requirements specified by the exam board. Thus, these findings indicated that the specs positively affected the test items, which can serve as validity evidence for the PTE General. Therefore, the insights gained here contribute to understanding the potential role of specs and how the quality of the specs and the items produced based on them can potentially support validity issues.

Conclusion and contributions of this research

The primary aim of this study was to examine the characteristics of reading test specs and IWGs, and the relationship between these two documents and items developed based on these documents. In particular, the study investigated the type of details in specs and IWGs that help item writers produce the intended items in terms of type and difficulty level. In addition, it explored how these two documents can facilitate the production of items at the intended level of difficulty. The secondary objective of this study was to study the use of the CEFR for specs and IWG development since the specs and IWGs investigated in this case study are for the reading part of the PTE General, which is linked to the CEFR. More specifically, this study examined how the CEFR language proficiency level descriptors lend themselves to producing precise specs and IWGs.

The study was carried out by adopting a triangulation approach, i.e., a combination of qualitative and quantitative methods for data collection and analysis were used to help ensure methodological validity. The qualitative approaches consisted of an analysis of the PTE General reading specs and reading IWGs, a focus group and individual interviews with PTE General item writers, and individual interviews with developers of the PTE General specs and IWGs. The quantitative methods concerned a two-step judgmental study by five language testing experts and an analysis of item statistics.

This study has led to empirically informed insights into the type of information item writers need in specs and IWGs to help them develop the intended items in terms of type and difficulty level. Furthermore, it serves as an original contribution to the field since there is a lack of research in language testing that has investigated the type of information perceived to be needed by item writers in specs and IWGs, the relationship between these two documents, and the link between the information provided and the items produced. In addition, this study investigated the IWGs that have hardly been covered in the language testing literature and have not been empirically investigated.

This study further shows no fixed conceptualization of specs and IWGs. Depending on the exam, the exam board, the assessment context, or the document users, the specs and IWGs could be developed as two separate documents, one comprehensive document, or two versions of the same document for different users. What is crucial is that the necessary content is included in the document(s). In this case study, the item writers and document developers considered the IWGs as the item writers' documents. In addition, the item writers preferred having one concise document that comprises the necessary information for item writing.

A unique contribution of this study is the mixed-method approach used to enhance the study's construct validity. Previous language testing research focused on test specs (Belyazid, 1996; Cho, 1995; Kennedy, 2007; Li, 2006) and used one or two data collection methods and analyses. In this study, qualitative and quantitative data collection and analysis approaches were triangulated to obtain insights from different sources and groups of participants. In addition, the mixed-method approach facilitated (1) confirming findings across methods and (2) supporting and strengthening findings from some methods with findings from other methods.

This study also provided insights into validation theory and practice by identifying the type of information the developers and item writers believed is needed in specs and IWGs to facilitate the development of the intended items, facilitating the production of tests that assess the intended constructs. In addition, it provided empirical insights into specs' role in the test development process and supported the AUA (Bachman & Palmer, 2010). Findings obtained from the two steps indicated that the type and level of details provided in the specs facilitated the development of items that were judged close to the intended CEFR levels and whose overall quality was judged positively. These findings can contribute to establishing backing to support the AUA of the test. This study, thus, contributed to the understanding of the potential role of specs in the AUA and how the quality of the specs and items produced based on it can provide backing to warrants and claims in the AUA.

This study has investigated areas (specs, IWGs, item writing) that have not been heavily researched or adequately covered in the language testing literature. In addition, the mixed-methods approach implemented in this study, whereby data and analyses were triangulated, is novel to research on specs and IWGs in language testing. Therefore, this study's findings will benefit academic researchers and specs and IWGs' developers, item writers, test developers, and consequently, test users, including decision-makers, teachers, parents, and test-takers.

Abbreviations

AUA	Assessment Use Argument
CEFR	Common European Framework of Reference
GD	General Description
IWGs	Item writer guidelines
MCQ	Multiple-choice questions
PA	Prompt Attributes
PTE General	Pearson Test of English General
RA	Response Attributes
RQ	Research questions
SI	Sample item
Specs	Test specifications
SS	Specification Supplement

Acknowledgements
Not applicable.

Authors' contributions

There is no other author other than myself as the research, data collection, data analysis, data interpretation, findings, and conclusion were solely done by me. The author read and approved the final manuscript.

Author's information

Dr Zahra Ali Al Lawati is the Director of the Centre for Preparatory Studies, Sultan Qaboos University, Muscat, Sultanate of Oman. She received her MA in Language Studies with focus on Language Testing from Lancaster University, UK, in 2002 and PhD in Applied Linguistics with focus on Language Testing from Lancaster University, UK, in 2014. Her research interests include language assessment, language education, critical thinking, and twenty-first century skills.

Funding

There was no funding body in the design of the study and collection, analysis, and interpretation of data and in writing of the manuscript.

Availability of data and materials

The datasets generated and/or analyzed during the current study are not publicly available due to the confidentiality of Pearson Edexcel. The data that support the findings of this study are available from Pearson Edexcel but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available.

Declarations

Competing interests

The author declares no competing interests.

Received: 15 May 2022 Accepted: 29 March 2023
Published online: 18 April 2023

References

Alderson JC, Cseresznyés M. (2005). Reading and use of English. In J. C. Alderson (Eds.), Into Europe: Prepare for modern English exams (Vol. 1, pp. 1–297). Available from http://www.lancs.ac.uk/fass/projects/examreform/into_europe/Reading_and_Use_of_English.pdf

Alderson, J. C. (2000). *Assessing reading*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732935

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S., & Tardieu, C. (2006). Analyzing reading and listening tests about the Common European Framework of Reference. The experience of the Dutch CEFR construct project. *Language Assessment Quarterly, 3*(1), 3–30. https://doi.org/10.1207/s15434311laq0301_2

Arhin, A. K., Essuman, J., & Arhin, E. (2021). Analysis of item writing flaws in a communications skills test in a Ghanaian University. *Afr J Teach Educ, 10*(2), 121–143. https://doi.org/10.21083/ajote.v10i2.6762

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford University Press.

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

Belyazid S. (1996). Task-based language test specifications designed for an adult TEFL context in Morocco. Unpublished MA thesis, University of Illinois at Urbana-Champaign, USA.

Cho D. (1995). The effect of specificity of language test specifications on item construction. Unpublished PhD thesis, University of Illinois at Urbana-Champaign, USA.

Davidson, F. (2012b). Test specifications and criterion-referenced assessment. In G. Fulcher & F. Davidson (Eds.), The Routledge handbook of language testing (pp. 197–207). New York: Routledge. https://doi.org/10.4324/9780203181287.ch13

Davidson, F. (2012). Releasability of language test specifications. *Japan Language Testing Association (JLTA) Journal, 15*, 1–23.

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. https://doi.org/10.1017/S0261444807004351

Davidson, F., & Lynch, B. K. (2002). *Test craft*. Yale University Press.

Fulcher, G. (2021a). Language Assessment Literacy in a Learning-Oriented Assessment Framework. In A. Gebril (Ed.), Learning-oriented assessment: Putting theory into practice (pp. 254–270). New York: Routledge. https://doi.org/10.4324/9781003014102

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Routledge.

Fulcher, G., Panahi, A., & Mohebbi, H. (2022). Glenn Fulcher's thirty-five years of contribution to language testing and assessment: a systematic review. *Language Teaching Research Quarterly, 29*, 20–56. https://doi.org/10.32038/ltrq.2022.29.03

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. https://doi.org/10.1177/0265532211413445

Gutiérrez Baffil, T. G., & Collada Peña, I. D. L. C. (2022). Assessing writing in English in Cuban higher education. *Transformación, 18*(1), 238–252.

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), 309–334. https://doi.org/10.1207/S15324818AME1503_5

Haladyna, T. M., & Rodriguez, M. C. (2021). Using full-information item analysis to improve item quality. *Educational Assessment, 26*(3), 198–211. https://doi.org/10.1287/ited.2022.0274

Hambleton, R. K., & Eignor, D. (1979). *A practitioner's guide to criterion-referenced test development, validation, and test score usage (Report No. 70)* (2nd ed.). University of Massachusetts.

Harsch, C., & Seyferth, S. (2019). Marrying achievement with proficiency in developing and validating a local CEFR-based writing checklist. *Assessing Writing, 43*, 10–43. https://doi.org/10.32038/ltrq.2021.26.02

Hughes, A. (1989). *Testing for language teachers*. Cambridge University Press.

Huhta, A., Luoma, S., Oscarson, M., Sajavaara, K., Takala, S., & Teasdale, A. (2002). A diagnostic language assessment system for adult learners. J. C. Alderson (Ed.), Common European Framework of Reference for Languages: Learning, teaching, assessment. Case studies (pp. 130–145). Council of Europe.

Jin Y. (2021). Test specifications. In Fulcher, G & Hardling, L (Eds.), The Routledge handbook of language testing (pp.271–288). Taylor & Frances.

Jones, N. (2002). Relating the ALTE framework to the Common European Framework of Reference. J. C. Alderson (Ed.), Common European Framework of Reference for Languages: Learning, teaching, assessment. Case studies (pp. 167–183). Council of Europe.

Kennedy, L. C. (2007). Expanding test specifications with rhetorical genre studies and activity theory analyses. Unpublished MA thesis, Carleton University, Ottawa, Ontario.

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*, 160–174.

Li, J. (2006). Introducing Audit Trails to the World of Language Testing. Unpublished MA thesis, University of Illinois at Urbana-Champaign, USA.

Norris, J. M., Brown, J. D., Hudson, T., & Yoshioka, J. (1998). Designing second language performance assessments. Honolulu: Second language teaching & curriculum Centre, University of Hawaii/University of Hawaii Press.

Osterlind, S. J. (1998). *Constructing test items: Multiple-choice, constructed-response, performance, and other formats* (2nd ed.). Kluwer Academic Publishers.

Pearson. (2012). Test Centre handbook Available from http://pearsonpte.com/TestCenters/Pages/Resources.aspx

Popham, W. J. (1978). *Criterion-referenced measurement*. Prentice-Hall.

Roid, G. H., & Haladyna, T. M. (1982). *A technology for test-item-writing*. Academic Press.

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

Salisbury, K. (2005). The edge of expertise: towards an understanding of listening test-item-writing as professional practice: unpublished PhD thesis, King's College, University of London. https://doi.org/10.1002/9781118784235.eelt0981

Shi, D. (2021). Item writing and item writers. In Fulcher, G & Hardling, L (Eds.), The Routledge handbook of language testing (pp.341–356). Taylor & Frances.

Stake, R. E. (2008). Qualitative case studies. In N. K. Denzin & Y. S. Lincoln (Eds.), *Strategies of qualitative inquiry* (pp. 119–149). Sage.

Tashakkori, A., & Teddlie, C. (2003). The past and future of mixed-methods research: From data triangulation to mixed model designs. In A. Tashakkori & C. Teddlie (Eds.), Handbook of mixed methods in social and behavioral research (pp. 671–702). Sage Publications. https://doi.org/10.4135/9781506335193

Tinkelman, S. N. (1971). Planning the objective test. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 46–80). American Council on Education.

Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. Language Testing, pp. 22, 281–300. https://doi.org/10.1191/0265532205lt309oa

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing*, 19(4), 453–476. https://doi.org/10.1191/0265532202lt240oa.

Little, T. D., Simpson, R. B., & O'Connor, P. (2002). Statistical methods for research in education and psychology (Third edition). Pearson Education.

Morgan, D.L. (1997). Focus Groups as Qualitative Research. 2nd Edition. Thousand Oaks: Sage.

Morrow, K. (Ed.). (2004). Insights from the Common European Framework. Oxford University Press.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.