A comparative study of reading sections of high-stakes teacher-made tests in Portugal

Estudo comparativo das secções de leitura de provas de alto impacto elaboradas por professores em Portugal

Estudio comparativo de las secciones de lectura de exámenes de alto nivel elaborados por profesores en Portugal

ABSTRACT

This study investigates the validity of high-stakes, teacher-made tests as parallel measures of achievement of identical curricular goals. Focusing on three reading sections from tests designed in different schools across Portugal, the research assessed a sample of 75 students representative of the intended population. Two complementary approaches were adopted: expert evaluations and psychometric analyses. Findings revealed discrepancies among experts regarding construct coverage and highlighted variations in item scope, format, subcomponents assessed, and comprehension demands across the sections. Technical shortcomings in instructions, item design, and scoring criteria were also identified. Results from a one-way repeated measures ANOVA demonstrated significant differences in mean performance across the three sections, with section two producing notably lower scores compared to sections one and three. These results suggest that the sections are not equivalent in construct or difficulty. The study reinforces concerns about teachers' assessment literacy, test quality, and fairness in assessment.

Keywords: high-stakes teacher-made tests; validity; reliability; EFL; secondary-level education; teachers' assessment literacy.

CADERNOS de LINGUAGEM & SOCIEDADE

Papers on Language and Society

Margarida Maria Pato
margarida.pato@study.beds.ac.uk
margarida_pato@hotmail.com
orcid.org/0009-0008-3945-1577
University of Bedfordshire—Centre for Research in English Language Learning and Assessment **(CRELLA),** Bedfordshire**,** United Kingdom

Instituto de Avaliação Educativa (IAVE), I.P., atual Instituto de Educação, Qualidade e Avaliação (EduQA), I.P., Lisboa, Portugal

Recebido em: 8 de setembro de 2025
Aceito em: 24 de fevereiro de 2026
DOI: 10.26512/les.v26i2.59640

ARTIGO

RESUMO

Este estudo analisa a validade de provas de alto impacto elaboradas por professores como medidas paralelas de objetivos curriculares comuns. A investigação concentrou-se em três seções de leitura aplicadas em diferentes escolas de Portugal, envolvendo uma amostra de 75 estudantes representativos da população-alvo. Dois enfoques complementares foram utilizados: avaliações de especialistas e análises psicométricas. Os resultados mostraram discrepâncias entre especialistas relativamente à cobertura do construto e evidenciaram variações no escopo e formato dos itens, nos subcomponentes avaliados e nas exigências de compreensão. Foram igualmente identificadas deficiências técnicas relacionadas com instruções, elaboração dos itens e critérios de correção. A análise de variância de medidas repetidas (ANOVA) revelou diferenças significativas no desempenho médio das três seções. Os resultados sugerem falta de equivalência em termos de construto e dificuldade, reforçando preocupações sobre a literacia de avaliação dos professores, a qualidade das provas e a equidade na avaliação.

Palavras-chave: provas de alto impacto elaboradas por professores; validade; fiabilidade; EFL; ensino secundário; literacia de avaliação dos professores.

RESUMEN

Este estudio examina la validez de pruebas de alto impacto elaboradas por docentes como medidas paralelas de objetivos curriculares comunes. La investigación se centró en tres secciones de lectura aplicadas en diferentes escuelas de Portugal a una muestra de 75 estudiantes representativos de la población objetivo. Se utilizaron dos enfoques complementarios: evaluaciones de expertos y análisis psicométricos. Los resultados mostraron discrepancias entre especialistas respecto a la cobertura del constructo, el alcance y formato de los ítems, los subcomponentes evaluados y las demandas de comprensión. Se identificaron deficiencias técnicas en las instrucciones, la construcción de ítems y los criterios de corrección. El análisis de varianza de medidas repetidas reveló diferencias significativas en el rendimiento medio de las tres secciones. Estos hallazgos sugieren inequivalencia en constructo y dificultad, lo que refuerza preocupaciones sobre la competencia evaluativa de los profesores, la calidad de las pruebas y la equidad en la evaluación.

Palabras clave: pruebas de alto impacto elaboradas por docentes; validez; fiabilidad; EFL (inglés como lengua extranjera); educación secundaria; competencia evaluativa de los profesores.

Como citar:

PATO, Margarida M., A comparative study of reading sections of high-stakes teacher-made tests in Portugal. **Cadernos de Linguagem e Sociedade**, Brasília, v. 26, n. 2, p. 291-313, jul./dez. 2025. DOI: 10.26512/les.v26i2.59640 Disponível em: . Acesso em: XXX.

Correspondência:

Margarida Maria Pato

margarida.pato@study.beds.ac.uk

INTRODUCTION

Portuguese teachers have as a requirement to design tests at the end of the school year. These tests provide students finishing an education cycle a second chance of attaining approval or, in the case of secondary students, a chance of improving their mark. The singularity of these tests (Testes de Equivalência à Frequência in Portuguese) is that they are made by teachers at schools, while being simultaneously high-stakes tests. The tests discussed here are implemented at the end of both middle and secondary school levels. Their key point is to certify students' achievement in a specific subject as indicated in the existing curricular documents. As certification tests, these are expected to "meet particularly high standards of comparability, consistency, fairness, and validity" (U.S. Congress, 1992, p. 12). It is, therefore, assumed they represent the same construct, test students at the same level, possess the same difficulty level and demand similar cognitive processes. However, this assumption has never been questioned nor verified and therefore never validated, even though, according to research, all interpretations of test scores "must be validated" (Solórzano, 2008, p. 285) as "[a] language test without validation research is like a police force without a court system: unfair and dangerous" (McNamara, 2007, p. 280).

The cornerstone of the research project underlying this paper, resulting from a MA in Language Testing programme at Lancaster University, was the analysis of a sample of three of these tests with the aim of providing an insight into teacher-designed tests in Portugal and into their soundness and the validity of the inferences they generate. The paper presents the results of the comparative study of post-course achievement tests in the subject of English[1] as a second language designed in different schools/geographical areas. Due to a need to limit the object of study, only the reading section of three tests was considered[2].

The study analysed the construct being tested and the validity and equivalence of the inferences drawn from students' scores, thus test use was addressed "within the scope of validation" (McNamara, 2007, p. 280). The study is thought to constitute an "instrumental use of research (…) that [may inform] decision-making" (Farley-Ripple *et al.*, 2018, p. 237) processes and support or not the use of teacher-made tests as high-stakes tests. Its results may be used to reflect on other high-stakes, teacher-made tests, both at middle school and secondary level.

The next sections focus on i) the literature review, which provides a theoretical background of analysis of the comparability study at hand; ii) methodological options, data collection procedures and instruments; iii) results of the research; iv) discussion; and v) conclusion.

[1] The teacher-made tests of English at secondary level with the purpose of certifying achievement were in the meantime replaced by the national exam. However, this did not happen to all the subjects.
[2] The terms 'reading section' and 'test' will be used interchangeably throughout this paper.

1. LITERATURE REVIEW

As this study explored various facets of high-stakes teacher-made tests, the literature review also focused on different areas, namely validity, construct definition, and teacher-made tests. A brief account of the most relevant aspects is provided below.

1.1 Validity and Construct Definition

As the aim of this study is grounded on a concern related to the equity of high-stakes teacher-made tests and of the decisions taken as a result of students' performance on these tests, this research aligns with Chapelle's (1999) assertion that "[v]alidation begins with a hypothesis about the appropriateness of testing outcomes (i.e., inferences and uses)", it may be assumed that this is, above all, a validity study.

The contemporary view of validity, as an "overarching concept" (Baker, 2013), extends to considerations such as test alignment, cognitive demands, and scoring validity. It has, therefore, come to include "an evaluation of the adequacy and appropriateness of the uses that are made of assessment results" (Miller *et al.*, 2009, p. 71). The achievement of validation thus depends on the development of an argument that intends to justify the inferences made based on a test (Chapelle, 1999). This argument is "centred on construct validity" (Luoma, 2001, p. 65) and these inferences, therefore, relate back to the specific construct being tested.

The definition of construct is, for the above-mentioned reason, the keystone of any assessment instrument or test validation process (Luoma, 2001) and all the aspects referred to above "contribute to the notion of [that] construct" (Baker, 2013, p. 7). They are building blocks that contribute, each in its own way, to the balance of the final construction of the instrument/test and "[d]eficit in any one raises questions as to the well-foundedness of any interpretation of test scores" (Weir, 2005, p.13). It is through the selection not only of the content, tasks and "construct-based scoring criteria" (Messick, 1995, p.746), but also of the "metacognitive strategies" demanded to put into practice the desired "knowledge in language use" (Bachman, 1991) that the construct is defined. The construct or trait to be assessed is therefore shaped by all the circumstances dictating and surrounding the testing instrument as well as by "the [cognitive] processing necessary for task completion" (Weir *et al.*, 2009, p. 163) in that test.

The impossibility of testing a complex trait such as "communicative competence" in a single test implies sampling of behaviour that may allow us to judge students' language ability. This is the first hurdle for those writing a test. This sampling should be aligned with the type of information the audience of the test's results requires (McNamara, 1996). However, sampling always implies a subjective interpretation, an "interpretative construct" (Davies, 2007, p. 77) based on the reference documents and on a personal conceptualisation of the trait being assessed, which, together with factors such as time, context and marking criteria, may lead to lack of reliability and to unwanted

variability (Harlen, 1994, as cited in Cohen *et al.*, 2007). According to Kane (2006), validity is, therefore, based "on the appropriateness and clarity of the target domain, the representativeness of the sample of observations, the appropriateness of the scoring rule and procedures" (pp. 150, 151).

The issue of "representativeness" is for this reason unavoidable when discussing validity (Messick, 1995). Evidence needs to be provided that the "content of the test is representative" (Bachman, 1991, p. 681) and that it provides "useful (...) information about the domain it is supposed to assess" (Lee & Klein, 2002): Does the content of the test represent the construct intended? Does it cover it appropriately both in depth and breadth? Does it activate the intended cognitive processes? Or does it, inadvertently test the ability to answer to specific item types, "memory, metalinguistic awareness, or test wiseness?" (Troike, 1983, p. 209). These questions must be in the mind of the test developer constantly. Likewise, it becomes essential to ask: What exactly do scores mean? What do they indicate about the construct? It is the appropriateness of the test and its scoring criteria in relation to an ability, a target population and a purpose that may or may not actually guarantee that validity is observed. From this perspective, once the mastering of the construct is (or should be) embodied in a score, the interpretation of scores becomes a crucial aspect of the development of a validity argument because "validity deals with the meaning of the measure" (Luoma, 2001, p. 61).

Questions raised about scores are therefore even more paramount because present perspectives on the construct of reading are embedded in a schema theory that puts the reader at the heart of the reading process, thus focusing on the "knowledge, experiences, values, and strategies" (Becker & Nekrasova-Beker, 2018, p. 297) they activate in the process. However, "explanations of complex cognitive processes [such as reading in a second language] have too many degrees of freedom" (Kintsch, 1998, p.1) and are, therefore, no easy task. Bachman (2007) warns that scores are sometimes related to latent traits that rely on subjective views and interpretations. He adds that while research has found that the same reading passage may be interpreted differently and pose different cognitive demands on each individual student, thus activating distinct processes in distinct students, score interpretation remains static, forcing the assumption that scores are to be interpreted in the same way for all students and that they represent a continuum in reading ability regardless of the items students may have got right or wrong or the weighting (a subjective decision of test designers) of the items. If this is so, the question of score interpretation and its "value implications (…) as a basis for action" (Messick, 1989, p. 5) are, in fact, "the key issues of test validity" (Messick, 1989, p. 5). The score is a quantitative reflection of performance, of "behavioral consistencies" (Chapelle *et al.*, 2011, p. 3), "a compound of contributions" (McNamara, 2011, p. 436) of all the factors and situations surrounding test administration and through which inferences about construct can be made (Chapelle *et al.*, 2011).

As a corollary of the most recent validity theory, the analyses of the tests under study relied on both experts' and statistical analyses in a pursuit of covering several facets and of "combining empirical evidence with theoretical statements" (Luoma, 2001, p. 69) which are, as seen, essential

to validity. The tests were, therefore, examined in terms of content, construct coverage, technical quality, results and consequences. An attempt was made to interweave theoretical and psychometric analyses in the study of the three sections of the tests. Even though "[v]alidity involves an overall evaluative [and thus subjective] judgment" (Miller *et al.*, 2009, p. 73, bold in the original), the inclusion and triangulation of these several perspectives will better inform the conclusions of the study and the possible generalisability of results, while simultaneously counteract any possible bias from the researcher.

1.2 Studies on teacher-made tests

Research consistently highlights issues in teacher-made assessments, questioning their reliability and validity (Alderson, 2001; Alderson 2005; Brookhart, 2001; Campbell, 2013; Malone, 2013; Marso & Pigge,1988; Marso & Pigge, 1991; Sanders & Vogel, 1993; Schneider *et al.*, 2013; Simsek, 2016). However, teachers' conceptualisation of their assessment literacy is not consensual. While some literature supports the fact that teachers themselves (Mertler & Campbell, 2005; Stiggins & Conklin, 1992; Vogt & Tsagari, 2014) seem to admit "they possess limited and superficial knowledge of testing: terminology, construction principles, use, and score interpretation" (Impara, *et al.*, 1991, p. 16); other studies conclude that teachers surveyed believe they possess "an adequate knowledge of testing" which they said had been learnt "through their experiences in the classroom" (Wise *et al.*, 1991, p.38).

Irrespective of teachers' beliefs, studies continue to warn about flaws in teachers' assessment skills. Black *et al.* (2010) reported that "teachers' [assessment] practices, even those under their control, lacked both the rigour and the comparability that would be required, both intra- and inter-school" (p.221). In a recent article, Simsek (2016) found evidence that items in teacher-made tests of different subject areas provided clues and hints that compromised the objective of those items and that the tests did not meet the requirements of "educational measurement" as "60% of the items had mistakes that needed to be corrected or improved" (p. 488). He concluded "that they [teachers] are not skilful enough in writing test items that measure student learning both in scope and depth" (p. 488). Aschbacher's (1999) study also revealed teachers' uncertainty in relation to the intended goals of tasks they had designed.

Literature also provides evidence that teachers test mostly low-order thinking skills and seldom cover a suitable range of cognitive levels (Broekkamp *et al.*, 2004; Brookhart, 2001; Campbell, 2013; Carter, 1984; Gullickson, 1993; Marso & Pigge, 1988; Oescher & Kirby, 1990; Simsek, 2016; Stiggins & Conklin, 1992), even though both content coverage and cognitive processes are of paramount importance to make valid judgments of students' achievement (Ferrara & Way, 2016). From the perspective of a cognitive framework for reading (Khalifa & Weir, 2009), this

implies teachers' tests do not cover the whole spectrum of the reading processes, which, in turn, compromises the social and cognitive dimensions that underlie construct concepts.

Also recurrent is the admonition related to teachers' general lack of knowledge in educational measurement (Mertler & Campbell, 2005; Simsek, 2016; Vogt & Tsagari, 2014; Wise *et al.*, 1991) and consequently their impossibility to critically analyse the results of their own tests, improve faulty items and ensure "tests serve the purposes for which they were designed" (Oescher & Kirby, 1990, p.4). As the recognition of the importance of skills in educational measurement seem to be interwoven with "skilfulness in assessment" (Alkharusi *et al.*, 2011, p. 115) itself, breaking this cycle may be a difficult task. Even more so because teachers seem to "simply emulate the testing structures and testing techniques they experienced during their own school years" (Ort, 1967, pp. 396-399).

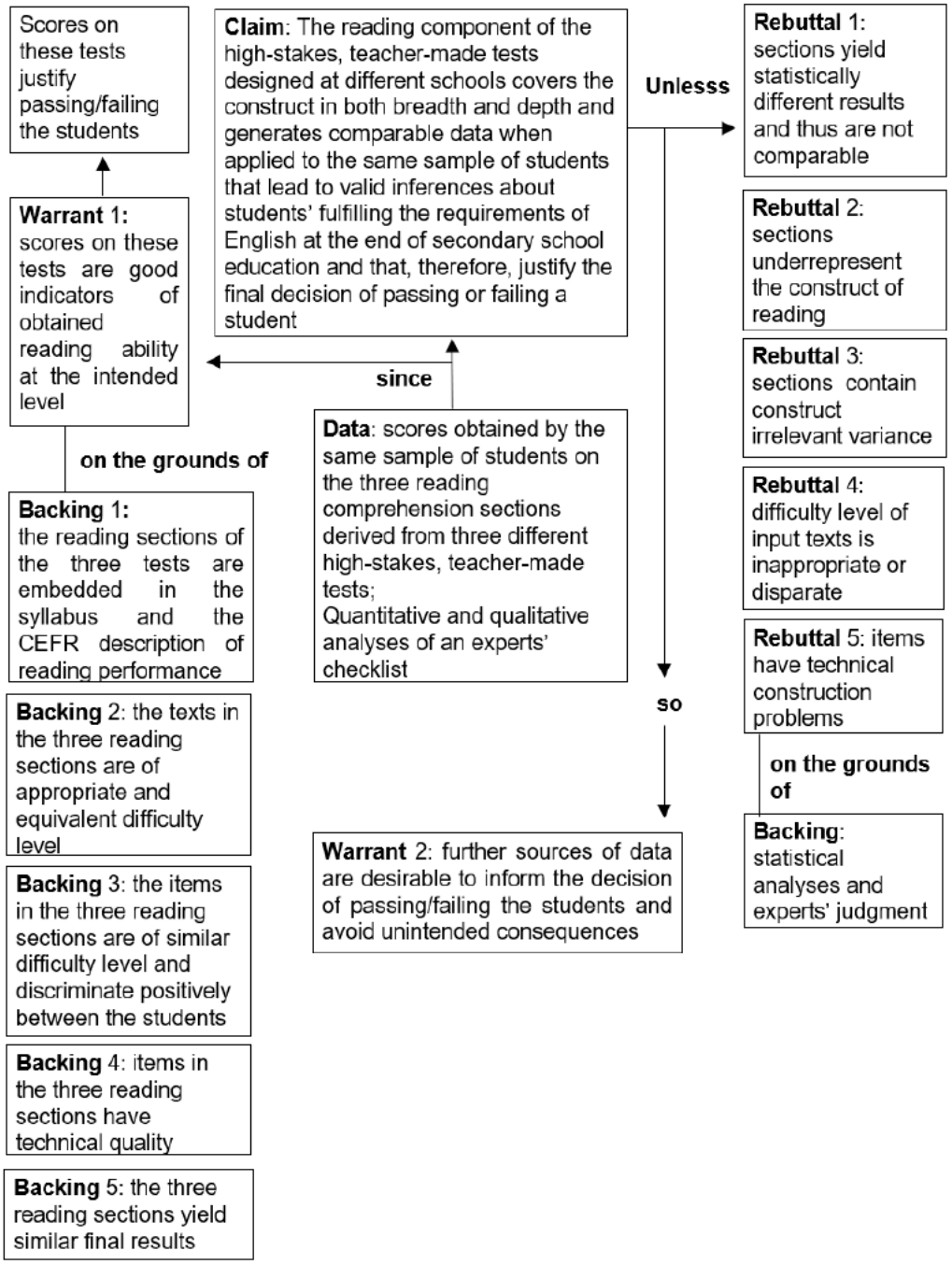
From the above, it becomes clear that to design a language test is a complex endeavour. It implies possessing both "the technical skills and knowledge to construct and analyse a test – the psychometric and statistical side to language testing – but it also requires a knowledge of what language is, what it means to 'know' a language" (Alderson, 2001, p.12). The question, then, is: are teachers well prepared for such an endeavour? Campbell's (2013) and Malone's (2013) answer is not reassuring: "[r]esearch [during the 90s and 2000s] (…) continued to document teachers' assessment skills as generally weak (…) [,] their tests poorly constructed and the results (…) frequently interpreted inaccurately" (Campbell, 2013, p. 72) and "many teachers (…) have a limited understanding of assessment fundamentals" (Malone, 2013, p.330).

Concerns over validity and teachers' lack of proficiency in constructing tests seem to be a constant in the assessment literature. Therefore, when considering the equivalence of high-stakes, teacher-made tests in Portugal, some questions need to be raised:

- Do the reading sections of the high-stakes school-made tests applied in different schools measure the same construct?
- Are they equivalent in their approach to the trait being measured?
- Do the tests yield equivalent psychometric results?
- Are the tests technically sound to allow high-stakes decisions based on them?
- To what extent are the tests fair to students who take them at different schools?
- To what extent can inferences drawn from test scores be valid?

Figure 1. depicts the model of validity argument, through which the present study attempts to answer these questions.

Figure 1 —The model of validity argument informing the present study



Source: Pato (2019)

2. METHODOLOGICAL OPTIONS, DATA COLLECTION PROCEDURES AND INSTRUMENTS

A mixed methods approach was adopted in this study, integrating both qualitative and quantitative techniques by means of five experts' views on the quality of the tests as measurement instruments of the construct of reading and statistical analyses of the test results when applied to a sample of students. Test selection followed a randomised approach, with tests sourced from eight

regions across Portugal. Some tests were excluded due to prior use in national exams or lack of scoring criteria.

Two instruments were applied: (i) a semi-structured checklist, completed by five testing experts, to evaluate the reading section of three different tests; and (ii) a reading test, composed of the three reading sections, administered to a sample of 75 11th-grade students, representative of the typical test-taking population. Before administration, the checklist was piloted by five testing experts and the wording problems were improved. The experts conducting the final analysis of the three sections were neither involved in the tests' design nor did they know the origin of the tests used in the study. Their professional background was different. While two of them were teachers of languages other than English, involved in text analyses from the technical point of view, the other three were involved in the design/technical analyses of tests while simultaneously working as English language teachers.

The checklist was designed to assess the technical characteristics of the reading sections, drawing from Khalifa and Weir's (2009) cognitive framework of reading and Kunnan and Carr's (2017) rating of item scope. The checklist focused on various aspects, including construct validity, cognitive processes, and technical test design features. Open-ended responses from the experts were qualitatively analysed, while closed-ended responses underwent statistical analysis. No reliability analyses of the checklist were carried out for several reasons. Namely, Cronbach's alpha values are "quite sensitive to the number of items in a scale" (Pallant, 2016, p. 95) and the checklist was composed of several subsets, mostly coded as multiple responses, thus representing very few items.

The reading sections scores were investigated by means of descriptive and inferential statistics to investigate whether the sections were equivalent in terms of content, construct and difficulty level, which "is important in understanding the meaning of [students'] achievement" (McNamara, 1996). Given that direct score comparisons across tests can be problematic (Allen & Yen, 1979), equating procedures (Kolen & Brennan, 2014) were considered but ultimately not applied as the three different sections might not comply with the requirements of equivalence for such a procedure. Instead, data distribution normality allowed for one-way repeated measures ANOVA to compare mean scores.

To control for test method effects, all students received the same three reading sections in varying orders. Data analysis was conducted using Coh-Metrix (v3.0), Excel (v14.4.2), and IBM SPSS (v24). Missing responses were initially coded as 99 for frequency checks but were later recoded as 0 to reflect real test-taking conditions and better reflect the tests' performance as administrated in Portuguese schools.

This methodological framework was believed to ensure a comprehensive evaluation of the technical quality, construct validity, and psychometric properties of the reading sections, aligning with best practices in language testing research.

The study adhered strictly to the ethical guidelines of the Lancaster University. All materials were anonymised prior to analysis, and authorisation for administering the test sections was obtained from the Pedagogical Council, informed consent was obtained form the participants. Particular attention was paid to protecting students' interests and well-being. Due to confidentiality constraints, the texts and test items are not publicly accessible.

3. RESULTS

The present study was developed to build a validity argument (Bachman, 2005; Chapelle *et al.*, 2011; Kane, 2011, 2013; Toulmin, 2003) of three reading sections of high-stakes, teacher-made tests that might justify or refute the use of these tests for their intended purpose. It attempted to either gather "evidence to justify interpretations based on performance" (Douglas, 2001) and consequently to justify the use of these tests to inform high-stakes decisions and thus claim students have achieved the intended level, or to find "alternative explanations, or counterclaims" that might "correspond to potential sources of invalidity" (Bachman, 2005, p. 10, italics in the original) and that might, therefore, challenge score interpretations and decisions emanating from administration of these tests.

The sections that follow give an account of the investigation of the core aspects underlying both warrants and rebuttals within the validity argument informing the present study.

3.1 The texts

This section will start with a brief comparison of the input texts of the three reading sections to investigate their appropriateness (Backing 1 and 2). It will then proceed to present the empirical results of the statistical analyses of the items and final scores and the analysis of the experts' answers to the checklist (Backing 3, 4 and 5).

As "the complexity of a text can make items designed to measure a student's ability (…) vary in difficulty" (Schneider *et al.*, 2013, p. 59), it seemed natural to start with the characteristics of the texts *per se*.

Data provided by the experts' checklist related to the appropriateness of text level, length and topic yielded the overall results in Table 1. Experts' answers were given in a Likert scale (1—Not at all—to 4—Definitely). Missing answers were not included in central tendency and dispersion analyses. These were only observed in one case in two of the reading sections for the variable of appropriateness of difficulty level.

Table 1 — Means and standard deviations by category (text difficulty level, length and topic)

Reading section	Difficulty level	Text length	Topic
One	3.80 (SD=0.447)	4 (SD=0.0)	4 (SD=0.0)
Two	3.50 (SD=0.77)	3(SD=1.225)	4 (SD=0.0)
Three	4 (SD=0.00)	4 (SD=0.00)	4 (SD=0.00)

From Table 1, according to the experts, it seems that section 2 is the least appropriate in terms of both difficulty level and text length. All the topics were deemed appropriate.

The three reading sections were also analysed by means of Coh-Metrix (v3.0) to check for similarity/dissimilarity regarding a set of parameters (Grassers *et al.* 2011) usually connected to readability difficulty. Data resulting from this analysis are presented in Table 2.

Table 2 — Results from the Coh-Metrix analysis of the three sections

	No paragraphs	No sentences	Sentence length mean	No Words in text	Type-token ration	No Sentence/ paragraph	Flesch reading ease	Flesch-Kincaid grade level	Coh-Metrix L2 readability
One	5	31	16.645	516	0.691	6.200 (SD 2.490)	59.910	9.038	11.109
Two	5	16	20.688	331	0.748	3.200 (SD 0.837)	33.218	13.766	7.609
Three	5	31	16.581	514	0.681	6.200 (SD 3.271)	65.559	7.609	20.551

Becker and Nekrasova-Beker (2018) used four features of input texts that seem to have a close relationship with difficulty. For the present analysis, three of those features will be utilised: i) general complexity—considered as the total number of words; ii) syntactic complexity—the mean length of sentence and clauses per sentence; and iii) lexical complexity—type-token ratio. The analysis of Table 2 focused on these aspects and on the readability formulas.

Even though the text from section two may seem an easier text as it is shorter (85 words and 83 words less than text one and text three, respectively), data resulting from the analysis of syntactic complexity and lexical complexity indicates that this section implies a higher complexity when compared to reading sections one and three. Considering that the higher the indices of the Flesch reading ease and the Coh-Metrix L2 readability are, the easier to read the text is, and that the higher the Flesch-Kincaid grade level index is, the less readable the text is (Kiselnikov *et al.*, 2020), it becomes evident that section three is the easiest to read and section two the hardest. In general, this is also aligned with the experts' analyses, considering text two was the one that resulted in more disparate judgments and lower means for appropriateness categories. There is a particularly indisputable difference between the sections in the Coh-Metrix L2 readability index. The results of the Coh-Metrix analysis seem to indicate a closer alignment between sections one and three in all the parameters except for the Coh-Metrix L2 readability index, which shows a higher discrepancy.

3.2 The structure

Each reading section was composed of a different number of items, which made “score comparisons more difficult” (LaFlair, *et al.*, 2017, p.128). There were also considerable differences in terms of item format. Table 3 illustrates these differences by presenting the structure of the reading comprehension section of the tests analysed and the respective total scores.

Table 3 — Structure of the three reading sections

Section	Item-format information	Item total score
1	Matching titles to paragraphs (5 paragraphs/6 titles, one not applied)	20
	Matching pronouns to the expressions they refer to	5
	Matching words or expressions from the text with synonyms (4 words or expressions/6 synonyms, two not applied	5x5=25
	5 multiple choice items	20
	Short answer: reporting a sentence from the text	5
	Gapped text (3 gaps/5 sentences)	
2	Short answer: identifying specific information	14
	Short answer: explaining the writer’s purpose of using a connector	8 2x10=20
	Short answer: completing sentences with information from 2 paragraphs	10
	Short answer: explaining an expression in context	5x2=10
	Short answer: identifying ideas replaced by anaphoric words in the text	3x6=18
	Short answer: replacing words/expression from the text with a synonym	
3	Matching information to paragraphs	24
	Short answer: completing sentences with information from the text	3x8=24 4x4=16
	Short answer: identifying synonyms in the text for words given	4x4=16
	Short answer: identifying words/ideas replaced by anaphoric words in the text	

Descriptive statistics were run to investigate item behaviour. The first overall observation of the data revealed that all selection items were answered by the students, whereas 91% of short answer questions presented missing data.

3.3 Items

Experts’ content analyses were first examined to investigate agreement. Experts were asked to i) identify construct coverage based on Khalifa & Weir’s framework (2009); ii) judge instruction clarity and item characteristics; iii) classify item scope (Kunnan & Carr, 2017); and iv) examine scoring criteria. Due to word limit constraints, only the most significant results pertaining to aspects i), ii), and iii) are reported (see Pato, 2019). A deeper analysis of results is carried out in relation to the statistical procedures carried out at both item and section levels.

3.3.1 Experts' judgments on construct coverage

There seems to be a wide range of opinions related to items and construct coverage subcomponents. Lack of agreement or dispersion of opinions seems more evident in the frequencies observed for reading for specific details and inferencing. The fact that some items tapped into several subcomponents, assessed the same type of skill or were identical and repetitive, or even that they tested grammar and, therefore, should have not been included in this section, were aspects further mentioned by the experts.

3.3.2 Experts' judgments on instruction clarity and item characteristics

When asked about clarity of instructions, experts identified several issues worth mentioning. They identified problems with instructions which included confusing and incoherent layout decisions; misleading indications and occasional mismatches between instructions and the way items were formulated (some instructions indicated responses needed to fit the text, but some of the options given were not grammatically accurate); some instructions were vague and lacking enough information to ensure students could perform accurately and appropriately according to what was expected in the scoring criteria.

Experts were also specifically asked about the items, focusing their analyses on coverage, clarity, avoidance of clues, construction and independence and variety of scope. The most striking data refer to the clarity and technical construction of items, particularly in sections one and two: in section one with a mean=2.80(SD=.837) and 2.40(SD.894), respectively; in section two 2.80(SD=1.095) and 2.20(SD=.837), respectively. In sections three, item clarity was within acceptable parameters with a mean=3.20(SD=.837), while item quality obtained values of 2.60 (SD=.894).

Problems with item construction identified by experts in one or more of these sections include: implausible distractors; "obviously wrong" answers; more than one possible correct answer; options with punctuation clues; similarity between distractors and the information in the text; size discrepancy and lack of parallelism of the options in multiple choice questions; use of very long stems and some negative-worded options in multiple choice questions; and decontextualized gap to fill in (first sentence of a paragraph). In their comments to section two, items were considered "too open" with many response options and their objective vague and confusing. When referring to section three, three experts considered the items "not really" well-constructed.

3.3.3 Experts' judgments on item scope

As far as overall variety of scope is concerned, there was a large discrepancy in experts' classification of the items. However, most items fell in the categories of very narrow, narrow and

moderate scopes. It is interesting to note that regarding section three, one of the experts referred to cognitive complexity of the items as being very low.

3.3.4 Statistical procedures

Bearing in mind that facility values and discrimination indices are crucial for analysing test items as they provide valuable insights into the quality and effectiveness of the items, these indices were calculated for the three sections. Scatterplots were drawn to better illustrate the relationship between the two variables in relation to the three sections. When analysing the scatterplots, it is important to bear in mind that more detailed information about students' performance comes from items "for which difficulty is targeted to provide maximal psychometric precision for key ranges of the score scale" (Ferrara & Way, 2016). These are the items that yield a 50-percent facility value, followed by those with 40 to 60 percent and 30 to 70 percent (Green, 2013).

It is worth mentioning that for section one only items 1., 3. and 6. are within the parameters of 40 to 60 percent. Items 1., 4.4., 4.5, 5., 6. and in special item 4.3 revealed very poor discrimination indices (below .20). Only items 3., 4.1. and 4.2. are within an acceptable value (.38, .31 and .30, respectively). In section two, most items were not within the range of facility values considered most informative. Only one item had a facility value of 54%. Four items had discrimination indices below .30. All other items discriminated positively between higher- and lower-proficiency students. However, only one item had simultaneously optimal facility and discrimination values. Section three seems to have been a difficult test for the students as only one item obtained a facility value above 50%. Two items obtained nearly optimal facility values (47% and 49%, respectively); however, their CITC indices were very low. In the case of one item, it was negative. Two items contributed negatively to the section's internal consistency.

Data from the reliability analysis of reading section one indicated a very low value with an overall Cronbach's Alpha of .358. This value would increase to .445, .363 and .390 if items 4.3., 6. and 5. were to be deleted, respectively. The reliability Cronbach's Alpha value for section two was .600 and for section three was .457.

Bearing in mind that .8 is the minimum desirable Cronbach's Alpha (Pallant, 2016), even more considering this is a high-stakes test, the values obtained in the present reliability analyses are an indicator that deeper analyses of the items are necessary.

For the above-mentioned reason an Inter-Item correlation matrix for each section was generated. The matrices provided evidence of very low internal consistency between the items in the three sections. Several values indicated a negative correlation The mean of the Inter-Item Correlations in section one was .072 with values ranging from -.241 to .338; In section two the mean of the Inter-Item correlations was .134 with values ranging from -.186 to .410; and in section three the Inter-Item Correlation mean was .106 with a minimum of -.199 and a maximum of .505. In this

section several negative and low correlation values were found, which may indicate items are tapping into different constructs.

The problems identified in the psychometric analyses of the items seem to conform with the experts' judgments. Apart from the lack of clarity mentioned above, the lack of technical quality may have had a decisive impact on the results of the psychometric analyses.

The results obtained by the administration of the three reading sections to the group of 75 students were analysed by means of descriptive statistics to investigate the measures of central tendency and dispersion. Moreover, to explore the "interpretation of individual scores" (Anastasi, 1976, p. 128) in a more suitable way, i.e. to analyse how "true" students' scores are and to "describe the consistency of" (Popham, 2008, p.38) their performance, the standard error of measurement for the three tests was calculated using the 14.4.2. version of excel. As this measure specifies the degree of error one needs to consider when making inferences (Miller *et al.*, 2009) about each individual score and the plausible scores students may obtain "as a result of the unreliability of the assessment" (Tighe *et al.*, 2010, p. 2), it was considered a fundamental piece of the sections' analysis. Table 4 presents a summary of the results.

Table 4 — Summary of the statistical procedures applied to the three tests

Test	Mean (SD)	Median	Mode	95% confidence interval	Range	Minimum	Maximum	Standard error of measurement
1	38.61 (13.39)	40	45	35.54-41.68	60	10	70	10.68
2	30.24 (15.5)	31	22, 24, 28, 34, 40	26.67-33.81	62	0	62	9.65
3	34.77 (13.8)	36	44	31.58-37.97	68	4	72	10.22

A Scatterplot matrix provided evidence of positive correlations between the three tests. Therefore, a Pearson's correlation was conducted to determine the relationship between the scores obtained by students on the three sections. The correlations obtained can be, according to Salkind (2008), considered moderate in the cases of tests 1 and 2 (Pearson's r= 0.539, n=75, $p < 0.05$) and 1 and 3 (Pearson's r= 0.469, n=75, $p < 0.05$) and strong for tests 2 and 3 (Pearson's r=0.710, n=75, $p < 0.05$). The coefficient of determination for tests 1 and 2 is, therefore, 29%, which means the tests share a variability of 29%, representing 29% of overlap between performances on both tests. As for test 1 and 3, there is 21.9% of shared variance between the two, whereas for tests 2 and 3 this shared variance presents a value of 50%.

To investigate if differences in students' performances on the three reading sections were due to chance or to real differences in the assessment instrument a one-way repeated measures

ANOVA was conducted to compare the final scores on the reading sections of the three different tests. There was a significant effect for the reading sections, Wilks' Lambda= .251, $F(2,72)= 14.439$, $p<0.05$, multivariate partial eta squared=.283. Post-hoc comparisons applying a Bonferroni adjustment indicated that the mean score for reading section one (M= 38.61, SD=13.39) was significantly different from the mean score for reading section two (M= 30.24, SD=15.5). Likewise, it suggested a significant difference between the mean score for section two (M= 30.24, SD= 15.5) and section three (M=34.77, SD=13.8).

4. DISCUSSION

Overall, the three input texts focus on topics that are aligned with the curricular documents and are grade appropriate. Most experts considered the texts adequate in terms of length; only text two yielded most disparate results. Readability formulas also corroborate the higher level of difficulty of text two. The data seem to suggest the input texts are not identical in terms of difficulty level, which strengthens rebuttal 4 of the validity argument. The lower mean obtained by students in section two supports this evidence. The difference between the three sections in terms of structure and item format may indicate different approaches to the construct of reading and represent "various degrees of structure or constraint imposed on the students' responses" (Messick, 1994).

A limitation of the construct was also identified. Experts seemed to acknowledge that the items do not tap into all of the subcomponents and that the majority of items test comprehension within a limited scope and at the level of low-order thinking skills that demand less depth and breadth of comprehension. Items seem to tap mainly into discrete points of reading comprehension such as vocabulary knowledge (more specifically synonymy), syntactic parsing and identifying specific information focusing on "local comprehension (…) [and] on the understanding of propositions within the sentence" (Weir *et al.*, 2009). The recurrent testing of synonymy with no constraints except parallelism to the meaning and form in the text and no distractors is nonetheless questionable owing to the use of all types of dictionaries. This limitation of the subcomponents of the construct of reading, thus constituting a counter-argument for warrant 1 of the validity argument. Moreover, considering construct validity within the scope of the three reading sections and the strength of the claim that these are representative of the construct intended, both experts' analyses and psychometric results on the correlation between the items seem to suggest there is an underrepresentation of the construct in terms of the subcomponents tested and of the breadth of text comprehension demands—which supports rebuttal 2 of the validity argument.

The lack of technical quality and clarity of the items in the three sections as referred to by the experts and supported by statistical analyses are a warning signal and seem to support rebuttal 5 of the validity argument, especially considering the high impact of these tests on students' lives. Overall, evidence was found that items had construction problems, which is aligned with the

statistical results obtained. This situation is consistent with the findings in literature dedicated to teacher-made tests. There seemed to be no “transparent relationship between test tasks and the [sub]skills, knowledge or abilities they represent” (Green & Weir, 2004, p. 475), which is evidence of a validity problem as also referred to by Kane (2006).

Even though achievement tests usually imply low discrimination, several items in these sections should be revised as they present extremely low discrimination indices. They seem to be either flawed or, as experts mentioned in the case of three items, to have more than one correct option. This suggests “quality control procedure[s]” such as “a final verification of the scoring key by content experts” (Downing, 2006, p.17) might have failed. The very low facility and CITC values of some items allied to their reliability indices indicate they do not make a positive contribution to the test’s internal consistency and may provide little or even misleading information about students’ reading ability. This lack of internal consistency between the items may also reflect construction problems or of a possible “muddying” of the construct by the scoring criteria. The fact that these may “not reflect the relevant knowledge and skills could lead to erroneous scores” (Xi, 2008, p. 183) and may result in construct-irrelevant variance supporting rebuttal 3 of the argument validity and which might, in the case of administering these tests to different groups of students, represent a lack of equity as students with similar degrees of reading proficiency might obtain very different scores.

Bearing in mind that the cut score at Portuguese schools is always half the total marks, and that the standard error of measurement for reading section one is 10.68, this means there is 95% chance that the true score of a student with a score of 40 on this test may lie somewhere between approximately 18.64 and 61.36 (Miller *et al.*, 2009). As for section two, even though the estimate of the standard error of measurement for this test is slightly lower (9.65) than for reading section one, it still represents a very high margin of error, particularly for a high-stakes test as it indicates the true score of a student with a mark of 40 may be between the lower limit of 20.7 and the upper limit of 59.3. The standard error of measurement for section three was 10.22. The true score of a student with a mark of 40 could be within a range of 19.56 and 60.44. Considering the three tests’ high standard error of measurement values, it becomes necessary to question the tests’ reliability as measures of a construct. If the scores obtained by students can indeed vary so significantly, the tests may indeed lack reliability as they were not free of errors of measurement. Had only the reading section of the tests been considered, it seems that most students would fail: 52 students in section one, 52 students in section two, and 42 students in section three. This emphasises the fact that teachers need to be aware of “the imprecision of the test scores an individual receives” (Popham, 2008, p. 39. Italics in the original) and of the need to consider them with extreme caution, especially taking into consideration that this is a high-stakes test.

These tests are grounded on the same syllabus, have similar test syllabi and are built for the same proficiency level. Nevertheless, the computed ANOVA indicated that the probability is less than 5% that the differences obtained in the means were due to chance. This means we can,

therefore, reject the null hypothesis consisting of that there were no differences in the means achieved by students in the three sections (Pallant, 2016). As “effect size is seen as much more important than significance” (Cohen *et al.*, 2007, p. 520), effect size was calculated to indicate to what extent the scores in the three sections differ. The effect size obtained was .283, (partial eta-squared =.283) suggesting a large effect (Pallant, 2016). Post-hoc comparisons using the Bonferroni test indicated that the mean score for section two (M=30.24, SD=15.5) was significantly different from the mean scores of the other sections. According to Coh-Metrix, the text in reading section number two was indeed less aligned with the other two. These results seem to support rebuttal 1 of the validity argument: the reading sections did not yield similar results, which seems to be an indication of a lack of psychometric equivalence between them.

5. CONCLUSION

The research briefly outlined in this paper has come to conclusions that reinforce the problems already identified in literature in relation to teacher-made tests: “They [teachers] lack expertise in test construction” (Brookhart, 2001, p. 11). In this case, these problems pose an even more considerable threat as a result of the “power” of these tests.

The analyses conducted here have raised both theoretical and psychometric questions, which emphasise the need for careful consideration of test results as the only measure for decision-making. Moreover, they have raised ethical questions connected to the uniformity of the tests applied in the numerous schools around the country, which is embodied in the question: do “all test takers have the same opportunity to demonstrate their abilities” (Kunnan, 2018, p. 76)?

Other issues emerging form this study can be embodied by the following questions: If teachers have issues identifying goals and cognitive processes as well as constructing items with technical quality, what does this tell us about the classroom tests used by teachers throughout the year and that are one of the major criteria for assessing students in Portugal? Moreover, what does this tell us about pre-service teacher training in Portugal?

It is therefore of paramount importance to investigate to what extent teacher training in Portugal is providing teachers with the necessary tools to assess their students reliably and with fairness. Likewise, it is of paramount importance to ensure teachers are given continuous and sound in-service training throughout their careers. Moreover, it is also essential that stakeholders possess assessment literacy to be able to monitor and take decisions informed by research. It remains to be said that the limitations of this study are inextricably connected to the generalisability of results.

From the literature review and from this study, it seems that there are still many dilemmas to be addressed, namely: how to ensure teachers possess the technical quality needed to construct testing instruments? How to reconcile the need for equivalent tests with the need to ensure they are locally relevant and appropriate? How to design local tests that respond to national demands and

requirements? How to ensure high-stakes, teacher-made tests do not have “an impact that goes beyond the measure they attempt to fulfil” (Shohamy, 1982)?

In 1983, Herman and Dorr-Bremme questioned, “How effective are teacher-generated tests in revealing the insufficiencies in individual students’ learning? How valid are they in terms of students’ achievement?” (Herman & Dorr-Bremme, 1983, p. 16). More than three decades have gone by and yet the questions we need to ask remain the same.

REFERENCES

ALDERSON, J. (2005). *Diagnosing foreign language proficiency. The interface Between learning and assessment*. London, UK: Continuum.

ALDERSON, J. (2001) The shape of things to come: will it be the normal distribution? In MILANOVIC, M., WEIR, C. & ASSOCIATION OF LANGUAGE TESTERS IN EUROPE (2004). *European language testing in a global context: Proceedings of the ALTE Barcelona conference*, July 2001 (Studies in language testing ; 18). Cambridge: Cambridge University Press, pp. 1-26.

ALKHARUSI, H., KAZEM, A. & AL-MUSAWAI, A. (2011). Knowledge, skills, and attitudes of preservice and inservice teachers in educational measurement. *Asia-Pacific Journal of Teacher Education,* 39(2), pp. 113-123.

ALLEN, M., & YEN, W. (1979). *Introduction to measurement theory*. Monterey, Calif: Brooks/Cole Pub. Co.

ANASTASI, A. (1976). *Psychological testing* (4th Ed.). New York: London: Macmillan; Collier Macmillan.

ASCHBACHER, P. (1999). *Developing Indicators of Classroom Practice to Monitor and Support School Reform*. CSE Technical report 513. National Center for Research on Evaluation. University of California, Los Angeles.

BACHMAN, L. (2005). Building and supporting a case for test use. *Language Assessment Quarterly: An International Journal*, 2(1), pp. 1-34, DOI:10.1207/s15434311laq0201_1

BACHMAN, L. (1991) What does language testing have to offer? *TESOL Quarterly*, 25(4), pp. 671-704.

BACHMAN, L. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In FOX, J. *et al.* (Eds). *Language Testing Reconsidered*, 3 (pp. 41-71). Ottawa: University of Ottawa Press.

BAKER, E. (2013). *The chimera of validity*. Teachers College Record, 115. Columbia University.

BECKER, A. & NEKRASOVA-BEKER, T. (2018) Investigating the Effect of Different Selected-response Item Formats for Reading Comprehension, *Educational Assessment*, 23:4, 296-317. DOI: 10.1080/10627197.2018.1517023

BLACK, P. (2010). Validity in teachers’ summative assessments, *Assessment in Education: Principles, Policy & Practice*, 17(2), pp. 215-232. DOI: 10.1080/09695941003696016.

BROEKKAMP, H., HOUT-WOLTERS, B., VAN DEN BERGH, H. & RIJLAARSDAM, G. (2004). Students' expectations about the processing demands of teacher-made tests. *Studies in Educational Evaluation*, 30, pp. 281-304.

BROOKHART, S. (2001). The “standards” and classroom assessment research. Paper presented at the annual meeting of the American Association of Colleges for Teacher Education, Dallas, TX. https://archive.org/details/ERIC_ED451189.

CAMPBELL, C. (2013). Research on teacher competency in classroom assessment. In MCMILLAN, J. (Ed.). *Sage Handbook of Research on Classroom Assessment*. Thousand Oakes, CA: SAGE Publications, pp. 71-84.

CARTER, K. (1984). Do teachers understand principles for writing tests? *Journal of Teacher Education*, 35, pp. 57–60. Doi:10.1177/002248718403500613.

CHAPELLE, C. (1999). Validity in language assessment. *Annual Review of Applied Linguistics* 19, pp. 254-272.

CHAPELLE, C., ENRIGHT, M. & JAMIESON, J. (2011). *Building a validity argument for the test of English as a foreign language* (ESL & applied linguistics professional series). New York; London: Routledge. Taylor & Francis e-Library.

COHEN, L., MANION, L. & MORRISON, K. (2007). *Research Methods in Education* (6th Ed.), Taylor & Francis Group. London: Routledge.

DAVIES, A. (2007). Assessing academic English language proficiency: 40+ years of U.K. language tests. In FOX, J. *et al.* (Eds). *Language Testing Reconsidered*, 4 (pp.73-86). Ottawa: University of Ottawa Press.

DOUGLAS, D. (2001). Performance consistency in second language acquisition and language testing research: a conceptual gap. *Second Language Research*, 17(4), pp. 442-456.

DOWNING, S. (2006). Twelve steps for effective test development. In DOWNING, S., & HALADYNA, T. (2006). *Handbook of Test Development*, 1 (pp. 3-25). Mahwah, N.J.: Lawrence Erlbaum Associates.

FARLEY-RIPPLE, E., MAY, H., KARPYN, A., TILLEY, K. & MCDONOUGH, K. (2018). Rethinking connections between research and practice in education: A conceptual framework. *Educational Researcher*, 47(4), pp. 235-245.

FERRARA, S. & WAY, D. (2016) Design and development of end-of-course tests for student assessment and teacher evaluation. In *NCME Applications of Educational Measurement and Assessment: Meeting the Challenges to Measurement in an Era of Accountability*, 5. NCME Book Series. New York: Routledge. Retrieved from https://www.book2look.com/embed/9781135040154

GREEN, A. & WEIR, C. (2004). Can Placement Tests Inform Instructional Decisions? *Language Testing*, 21(4), pp. 467-494.

GREEN, R. (2013). *Statistical Analyses for Language Testers*. New York: Palgrave Macmillan.

GULLICKSON, A. (1993). Matching measurement instruction to classroom-based evaluation: perceived discrepancies, needs, and challenges. In WISE, S. (Ed.), *Teacher Training in Measurement and Assessment Skills*, 3. Lincoln, NE: Buros Institute of Mental Measurements, University of Nebraska–Lincoln. http://digitalcommons.unl.edu/burosteachertraining/3

HERMAN, J. & DORR-BREMME, D. (1983). In HATHAWAY, W., *Testing in the schools (New directions for testing and measurement*; 19). San Francisco: Jossey-Bass.

IMPARA, J., DIVINE, K., BRUCE, F., LIVERMAN, M. & GAY, A. (1991) Does Interpretive Test Score Information Help Teachers? *Educational Measurement: Issues and Practices*, 10(4), pp. 16-18.

KANE, M. (2006) Content-related validity evidence in test development. In DOWNING, S., & HALADYNA, T. (2006). *Handbook of Test Development*, 7 (pp.131-153). Mahwah, N.J.: Lawrence Erlbaum Associates.

KANE, M. (2011). Validating score interpretations and uses: Messick Lecture, Language Testing Research Colloquium, Cambridge, April 2010. *Language Testing*, 29(1), pp. 3-17.

KANE, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1-73.

KHALIFA, H., & WEIR, C. (2009). *Examining Reading: Research and practice in assessing second language reading* (Studies in language Testing, 29). Cambridge: Cambridge University Press.

KINTSCH, W. (1998). *Comprehension: A paradigm for cognition*. New York, NY, US: Cambridge University Press.

KISELNIKOV, A., VAKHITOVA, D., KAZYMOV, T. (2020). Coh-metrix readability formulas for an academic text analysis, *IOP Conference Series: Materials Science and Engineering*, Volume 890, International Scientific Conference on Socio-Technical Construction and Civil Engineering (STCCE - 2020) 29 April - 15 May 2020, Kazan, Russian Federation. **DOI** 10.1088/1757-899X/890/1/012207

KOLEN, M. & BRENNAN, R. (2014). *Test Equating, Scaling, and Linking: Methods and practices* (3rd ed.). Statistics for social and public policy. New York: Springer.

KUNNAN, A. & CARR, N. (2017). A comparability study between the general English proficiency test-advanced and the Internet-based test of English as a foreign language. *Language Testing in Asia*, 7(17), pp. 1-16.

KUNNAN, A. (2018). *Evaluating Language Assessments* (New perspectives on Language Assessment Series). New York: Routledge, https://doi.org/10.4324/9780203803554

LAFLAIR, G., ISBELL, D., MAY, L., ARVIZU, M. & JAMIESON, J. (2017). Equating in small-scale language testing programs. *Language Testing*, 34(1), pp. 127-144.

LEE, V. & KLEIN, S. (2002). Technical criteria for evaluating tests. In Hamilton, L. *et al.* (Eds). *Making Sense of Test-Based Accountability in Education*. Santa Monica, CA: RAND Corporation. https://www.rand.org/pubs/monograph_reports/MR1554.html.

LUOMA, S. (2001). What does your test measure? Construct definition in language test development and validation. Retrieved from http://www.solki.jyu.fi/vanhat/Luoma_Sari_2001_PhD_manuscript1.pdf

MALONE, M. (2013). The essentials of assessment literacy: Contrasts between testers and users. *Language Testing*, 30(3), pp. 329-344.

MARSO, R. & PIGGE, F. (1988). An analysis of teacher-Made tests: testing practices, cognitive demands, and item construction errors. Paper presented at the annual meeting of the National Council on Measurement in Education. New Orleans, Louisiana.

MARSO, R. & PIGGE, F. (1991) An analysis of teacher-made tests: item types, cognitive demands, and item construction errors. *Contemporary educational psychology*, 16, pp. 279-286.

MCNAMARA, T. (1996). *Measuring second language performance*. London/NY: Longman.

MCNAMARA, T. (2007). Assessment in foreign language education: The struggles over constructs. *The Modern Journal*, 91(2), pp. 280-282.

MCNAMARA, T. (2011). Applied Linguistics and Measurement: A Dialogue. *Language Testing*, 28(4), pp.435-440.

MERTLER, C. & CAMPBELL, C. (2005). Measuring teachers' knowledge & application of classroom assessment concepts: development of the Assessment Literacy Inventory. Paper presented at the annual meeting of the American Educational Research Association,Montréal, Quebec, Canada.

MESSICK, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, 18(2), pp. 5-11.

MESSICK, S. (1994). The Interplay of Evidence and Consequences in the Validation of Performance Assessments. *Educational Researcher*, 23(2), 13-23. Retrieved from http://www.jstor.org.ezproxy.lancs.ac.uk/stable/1176219.

MESSICK, S. (1995). Validity of psychological assessment. Validation from inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist*, 50(9), pp. 741-749.

MILLER, M., LINN, R. & GRONLUND, N. (2009). *Measurement and assessment in teaching*, 10th Ed. Merril/Pearson. New Jersey.

OESCHER, J. & KIRBY, P. (1990). Assessing Teacher-made tests in secondary Math and Science classrooms. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, MA. ERIC Document Reproduction Service No. 322 169.

ORT, V. (1967). Teacher-made tests. *The Clearing House*, 41(7), pp. 396-399.

PALLANT, J. (2016) *SPSS Survival Manual* (6th ed.). NY: Open University Press/McGraw-Hill Education.

PATO, M. (2019). *A comparative study of the reading section of high-stakes teacher-made tests in Portugal* [Unpublished Master's thesis]. Lancaster University.

POPHAM, W. (2008). *Classroom assessment: what teachers need to know*, 5th ed. Boston: Pearson/Allyn &Bacon.

SANDERS, J. & VOGEL, S., "3. The Development of Standards for Teacher Competence in Educational Assessment of Students" (1993). *Teacher Training in Measurement and Assessment Skills*. 5.http://digitalcommons.unl.edu/burosteachertraining/5.

SCHNEIDER, M., EGAN, K. & JULIAN, M. (2013). Classroom Assessment in the Context of High-Stakes Testing, In MCMILLAN, S. (Eds), *SAGE Handbook of Research on Classroom Assessment*, 4, pp. 55-70.

SIMSEK, A. (2016). A comparative analysis of common mistakes in achievement tests prepared by school teachers and corporate trainers. *European Journal of Science and Mathematics Education*, 4(4), pp. 477-489.

SHOHAMY, E. (1982). Affective considerations in language testing. *The Modern Language Journal*, 66(1), pp. 13-17.

SOLÓRZANO, R. (2008). High stakes testing: issues, implications, and remedies for English language learners. *Review of Educational Research*, 78(2), pp. 260-329.

STIGGINS, R. & CONKLIN, N. (1992). *In teachers' hands: investigating the practices of classroom assessment*. Albany, NY: State University of New York Press.

TIGHE, J., MCMANUS, I., DEWHURST, N., CHIS, L. & MUCKLOW, J. (2010). The standard error of measurement is a more appropriate measure of quality for postgraduate medical assessments than is reliability: an analysis of MRCP (UK) examinations. *BMC medical education*, 10(40). Doi:10.1186/1472-6920-10-40.

TOULMIN, S. (2003). *The uses of argument* (Updated ed.). Cambridge, England: Cambridge University Press.

TROIKE, R. (1983). Can language be tested? *The Journal of Education*, 165(2), pp. 209-216.

U.S. CONGRESS (1992). *Testing in American Schools: Asking the Right Questions*, OTA-SET-519. Congress of the U.S., Washington, DC. Office of Technology Assessment.

VOGT, K. & TSAGARI, D. (2014). Assessment Literacy of Foreign Language Teachers: Findings of a European Study. *Language Assessment Quarterly* 11(4), pp. 374-402. DOI: 10.1080/15434303.2014.960046.

WEIR, C. (2005). *Language testing and validation [electronic resource]: An evidence-based approach* (Research and practice in applied linguistics). Basingstoke: Palgrave Macmillan.

WEIR, C., HAWKEY, R., GREEN, A. & DEVI S. (2009). *The cognitive processes underlying the academic reading construct as measured by IELTS in Research Reports*, 9, pp. 157-189, British Council/IDP Australia.

WISE, S., LUKIN, L. & ROOS, L. (1991). Teacher beliefs about training and measurement. *Journal of Teacher Education*, 42(1), pp.37-42.

A AUTORA

Margarida Maria Pato

Texto