Journal of Educational and Psychological Studies - Sultan Qaboos University (Pages 642-655) Vol.14 Issue 4, 2020
DOI: http://dx.doi.org/10.24200/jeps.vo14iss4pp642-655

Establishing the Validity of the Reading Questions in a Centralized Test Using Weir Socio-Cognitive Framework

Sheikha Al-Buraiki*
United Arab Emirates University, United Arab Emirates

Received: 1/7/2020 Accepted: 1/9/2020

Abstract: Establishing test validity is among the highly significant issues in language assessment which can be achieved by employing well-established validity frameworks. Adopting validity frameworks could generate valid and reliable tests that inform more systematic decisions. Using Weir's socio-cognitive framework (2005), this paper aims to highlight the validation process of the reading questions in the General Education Diploma of English Language Test (GEDELT) of 2016/2017 in Oman. Findings revealed that context-validity is inadequately satisfied due to the test response format, absence of allotted time for each question and the exhaustion that the test takers may experience due to the length of the test. Theory-based validity witnesses strengths from utilizing a large number of texts and a weakness from overemphasis on the skill of scanning to locate specific information. Scoring-validity is considered high since types of task response, marking guides and electronic marking reduce markers' subjectivity and minimize human error. The study draws its conclusions in light of the findings of test validity.

Keywords: Weir framework, socio-cognitive, validation, test validity, reading.

تحديد صدق أسئلة القراءة في اختبار مركزي باستخدام إطار العمل الاجتماعي المعرفي لوير

شيخة البريكي*
جامعة الإمارات العربية المتحدة، الإمارات العربية المتحدة

مستخلص: يعتبر تحديد صدق الاختبار من بين الأمور الأكثر أهمية في إطار التقييم اللغوي، والتي يمكن تحقيقها باستخدام أطر الصدق محكمة الصياغة ؛حيث ينجم عن أطر الصدق المذكورة اختبارات تتسم بالصدق والموثوقية من شأنها أن تؤدي إلى اتخاذ قرارات تتسم بقدر أكبر من المنهجية، لذلك تسلط هذه الورقة البحثية الضوء على عملية التحقق من صدق أسئلة القراءة في اختبار دبلوم التعليم العام في مادة اللغة الإنجليزية للعام الأكاديمي ٢٠١٧/٢٠١٦ في سلطنة عُمان وذلك باستخدام الإطار الاجتماعي المعرفي لـوير ٢٠٠٥، حيث كشفت النتائج أن صدق السياق غير كافي، ويعزى ذلك إلى نوع السؤال و طريقة الإجابة عليه وغياب تخصيص وقت محدد لكل سؤال وإلى الإجهاد الذي قد يتعرض له الممتحنون بسبب طول فترة الاختبار. وقد أظهرت النتائج أيضا أن الصدق المستند إلى النظرية اتسم بمواطن قوة تعود إلى استخدام عدد كبير من النصوص، إضافة إلى موطن ضعف يعزى إلى التركيز الزائد على مهارة المسح لتحديد معلومات بعينها داخل النص. كما يعتبر صدق التصحيح مرتفع نظراً لأن أنواع الاستجابة للمهام و توفر نماذج إجابة والتصحيح الإلكتروني من شأنها أن تقلل من تحييز التصحيح وتحد من الخطأ البشري، لذلك تضع هذه الدراسة نتائجها في ظل نتائج صدق الاختبار.

الكلمات المفتاحية: إطار وير، المعرفي- الاجتماعي، التحقق، صدق الاختبار، القراءة.

*201890041@uaeu.ac.ae

Various assessment methods are being used for a multiplicity of purposes, one of which is determining the eligibility for admission in high educational institutions. Grade twelve students, in the Sultanate of Oman, sit for final examinations for all the subjects they take. All the test subjects are centrally prepared by the assessment and supervision department in the Ministry of Education (MoE). For all the school subjects, 30% of the marks are allocated for continuous assessment and 70% for the final examination. Grade twelve marks the end of high school and great emphasis is placed on the final examinations because they determine students' future tertiary education tracks and majors, and subsequently determine their future employment. In the arena of testing and assessment, using a validity framework can guarantee valid and consistently reliable test scores and thus more reliable decisions. Therefore, establishing the validity of grade twelve English language test (ELT) is of an urgent need to ensure trustworthy and fair decisions, and to inform policy making in education.

Theoretical Framework

Carrell, Pharis and Liberto (1989) pointed out that reading is viewed as a key to success in higher educational institutions. Therefore, in order to build decisions on students' reading proficiency, designing valid and reliable reading tests represents a necessity. Test validity has been defined by Plake and Wise (2014) as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (p. 11). Various approaches have been proposed vis-à-vis reading assessment; however, this review will only shed light on Weir's socio-cognitive approach to test validation.

This research paper adopts Weir's (2005) socio-cognitive model for test validation. Weir's framework tests the cognitive dimension of the test takers and the social dimension of the test. In other words, the framework focuses on the abilities to be tested that can be referred to as mental constructs and it views language as a social rather than linguistic phenomenon. Weir provided frameworks for four language skills namely reading, listening, speaking and writing. Figure 1 provides a visual representation of Weir's validation framework for reading.

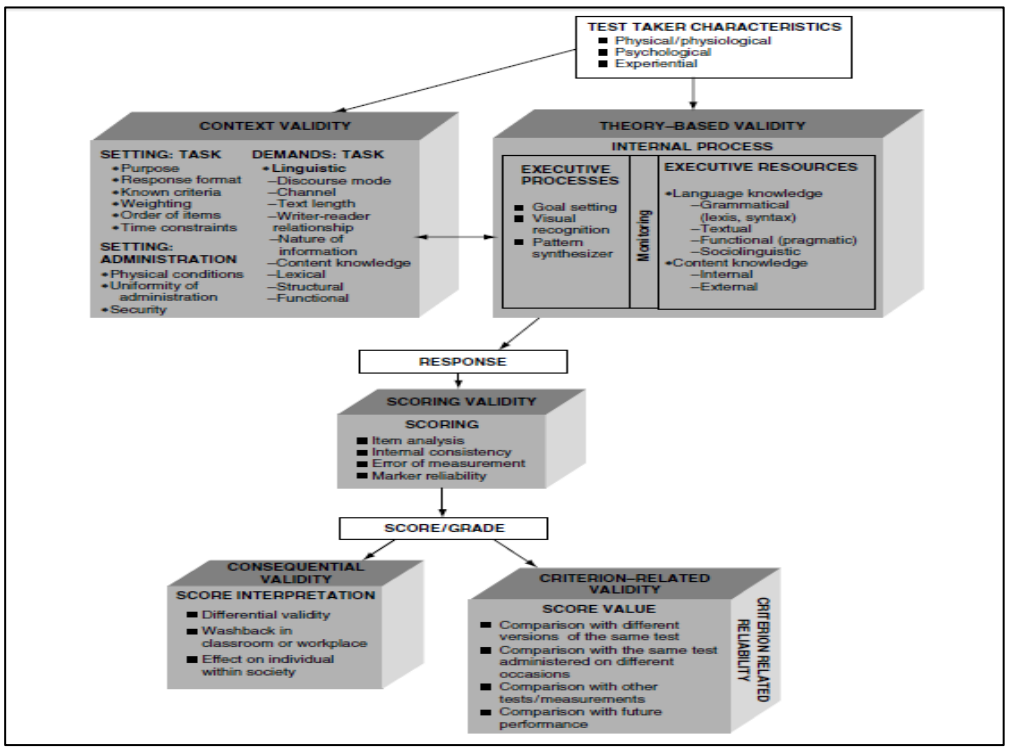


Figure 1. **The socio-cognitive framework for validating reading tests for Weir (2005)**

Weir's model encompasses two main stages for test validation: before the test which includes context and theory-based validity, and after the test which involves the components of scoring, consequential and criterion-related validity. Following is a detailed description of the framework.

Context validity refers to "the extent to which the test appropriately samples from the domain of knowledge or skills relevant to performance in the criterion" (McNamara, 2000, p. 132). The context, as Weir (1993) indicated, must be appropriate and acceptable to the test takers and test writers alike in order to assess specific language abilities. Context validity for reading tests—includes: task setting, task demands and setting administration, each is further divided into sub components (see figure 1).

Theory-based validity directly relates to the mental or cognitive processes associated with the acquisition of linguistic knowledge. According to Weir (2005), theory-based validity involves executive processes which contain goal setting, monitoring, visual recognition, and pattern synthesizer, executive resources that entail language knowledge (grammatical knowledge: lexical and syntax), textual knowledge, functional (pragmatic) knowledge, and sociolinguistics knowledge, and content knowledge consists of internal (background knowledge) and external (task knowledge). As the framework displays, there is an interactive relation between context and theory-based validity with scoring validity.

Scoring validity refers to all test aspects that influence scores' reliability which include item analysis, internal consistency, error of measurement and marker reliability. Item analysis concerns with analyzing item difficulty using statistical measurements to provide additional information about test takers and their abilities. Internal consistency can be computed statistically to gain information about test homogeneity. Error of measurement shows the difference(s) between the observed score and the corresponding true score or proficiency. Marker reliability primarily relates to the scoring process of tests and is usually influenced by the test type (objective or subjective), number of raters, and the method of scoring (manually or mechanically) (Weir, 2005).

Consequential validity refers to the impact of the test scores and interpretation at micro and macro levels: on the test takers, educational system and the society as a whole (Weir, 2005). Impact or washback, which is defined as the test effect on teaching, learning and testing (Fulcher, 2010), of the test can be either positive or negative. The aspects of consequential validity presented in the framework deals with the effect of tests in three areas: differential validity, washback in classroom or workplace and the effect on the individual within society.

The relationship between test scores and other external measurements that assess the same ability is what criterion-based validity is about (Weir, 2005). The framework illustrates the external measurements (elements) that can be incorporated with the scores or test values to examine criterion-related validity. These elements are: comparison with different versions of the same test (parallel or equivalent forms), comparison with the same test administered on different occasions, comparison with other tests/ measurements and comparison with future performance.

Talking about assessment-related topics and concerns about examinations, few studies were conducted in high school. To my best knowledge, there is no study in the Omani context that evaluated the validity of the GEDELT from a socio-cognitive perspective. However, blaming the assessment domain in the Omani educational system is an issue that has been raised by several researchers. For example, Al-Issa and Al-Bulushi (2012) mentioned the dominance of exam-based assessment in the Omani ELT educational system.

A number of studies shed light on the impact of exams on teaching and learning. For instance, Mohammed (2019) explored the reasons why learning English in school for twelve years remains inadequate to walk students through tertiary education without going through a foundation year. Mohammed concluded that teachers' concerns about preparing students to meet

final exam's requirement made their instruction teacher-centered which subsequently resulted in students' poor proficiency level in English language. The assessment system of grade twelve, where the final centralized test constitutes 70%, resulted in teachers' "teaching to the test" (MoE and World Bank, 2012, p. 31). Studying the washback effect among grade twelve teachers and students in Oman, teachers in Al-Lawati's (2002) research reported their shift in focus to exam-based instruction particularly when end of semester exams were approaching. In the same study, grade twelve students, noted that they geared their study to be exam oriented maneuvering their focus to the language components included in the final test.

Al-Mekhlafi, Al-Mamari and Al-Barwani (2019) assessed the presence of communicative competence features in grade ten ELT in Oman. Findings revealed a discrepancy between the test specification and the actual test content, and that the communicative competence was not fully addressed in the test. Reading-related findings unfolded the assessment of other language components in reading questions other than reading per se, e.g. vocabulary knowledge. The researchers recommended the alignment of testing with the mode of instruction adopted in schools. Additionally, the educational scene in Oman presents a need to further carry out studies in the area of assessment particularly in English language subject matter. Some previously conducted studies, e.g. Al-Kharusi (2011), recommended further studies analyzing teachers' assessment practices in the Omani context.

Problem Statement

The accurate validation of ELTs has increasingly become a major demand in L2 contexts. A validation process empirically verifies that a test task measures what the task aims to measure and that the inferences made based on the test scores are valid (e.g. Bachman, 1990; Weir, 2005). The need for validating local tests has been stressed by a number of researchers. For example, Al-Ismaili (2015) stated that "many local tests do not have a rigorous system of validation…[and thus] the need for such a system of evidence-based validation studies has been officially recognized for second language testing in Oman" (p. 6).

The GEDELT in Oman has been used to assess students in their English subject for many years. Very few students manage to score As and Bs in English test with the majority falling in the C and D categories. According to the results of the ELT of the school year 2016/2017, 14% of the total passers in the English subject scored A, 16% scored B, 30% scored C and 40% scored D (Higher Education Admission Center, 2019). With many students scoring low in English language coupled with going through foundation programs in subsequent tertiary institutions, a need arises to look closely into the test and examine its validity and reliability if major curriculum reforms are to take place. The main purpose of this study is to validate the reading component of the GEDELT as it accounts for 30% of the total mark in the English subject. Thus, empirical-based evidence is highly needed (e.g. Bachman, 2000; Weir, 2005) to inform teaching and learning practices and to urge decision makers to reconsider the reconceptualization of assessment practices and curriculum development at large.

Based on the above discussion, and the increased needs to establish the validity and reliability of centralized high-stake exams, the present study seeks to answer the following research question: How valid is the General Education Diploma of English Language Test (GEDELT) in Oman in light of Weir's socio-cognitive framework?

Methodology

Research Design and Data Collection

The present research employs Weir's (2005) socio-cognitive framework which is a comprehensive checklist approach for collecting detailed evidences that fulfill the components of the framework. A triangulation method was utilized for data collection through which two sources were used; namely a checklist, and document analysis. I have developed a check-

list including all the detailed components of Weir's framework (see table A1).

Additionally, I relied on analysis of several official documents. Document analysis is a systematic research procedure used to review and evaluate document materials (Bowen, 2009) in order to elicit meaning, to empirically develop knowledge and to gain deeper understanding (Rapley, 2007). These documents include the Student's Assessment Handbook (SAH) for English, Grades 11 and 12 (MoE, 2016), the English Language Teaching Curriculum Framework (MoE, 2012), a document of test regulations and administration (MoE, 2015), and a document released by the Higher Education Admission Center in 2019. Table B2 illustrates how each one of these documents have been utilized to associate relevant evidence with test items.

Trustworthiness and Credibility

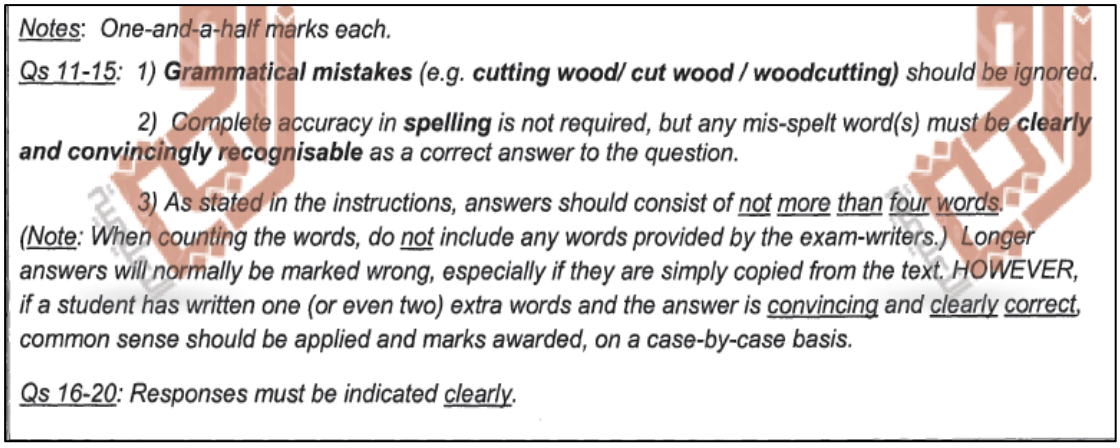
Unlike quantitative research studies where the researcher seeks to uncover the truth and thus to establish the validity and reliability of the data collection tools and data analysis, qualitative studies shift the focus to trustworthiness and credibility in order to minimize researcher's bias and subjectivity (Angen, 2000). With respect to trustworthiness, I used member checks, rich and thick description and investigator expertise (see Creswell, 2007). Member checks method was achieved via getting an English language supervisor, who has a large experience in supervision, classroom observations, and assessment, to review the analysis procedures and the findings.

Table B2

Documents utilized to link evidence with the test items

validity domain	framework check-list	documents used	usefulness of the document(s)
test takers characteristics	√	regulations of managing general education diploma exams (2015)	Items in the document pinpointed the accommodation provided for the test takers physical conditions.
context	√	• student assessment handbook • ELT curriculum framework • regulations of managing general education diploma exams (2015)	• The document was useful to find evidence for test specifications. • The document was useful to locate information on the genres of texts in grade 12 syllabus. • The document was consulted to find evidence that cater for selection of test sites, the total number of candidates admitted in each site, number of test takers per each classroom, number of invigilators, general instructions for test takers, invigilators, and site administration.
theory-based	√		
scoring	√		
criterion-related			
consequential	√	number of general education diploma students who passed English language subject and got admitted into higher education in 2017/2018 (2019)	The document was useful to outline the general grades of the students and how the score in English determined students' admission to higher educational institutions.

Note. I tabulated the documents used as data collection as a method to examine test validity



Notes: One-and-a-half marks each.

Qs 11-15: 1) **Grammatical mistakes** (e.g. **cutting wood/ cut wood / woodcutting**) should be ignored.

2) Complete accuracy in **spelling** is not required, but any mis-spelt word(s) must be **clearly and convincingly recognisable** as a correct answer to the question.

3) As stated in the instructions, answers should consist of not more than four words. (Note: When counting the words, do not include any words provided by the exam-writers.) Longer answers will normally be marked wrong, especially if they are simply copied from the text. HOWEVER, if a student has written one (or even two) extra words and the answer is convincing and clearly correct, common sense should be applied and marks awarded, on a case-by-case basis.

Qs 16-20: Responses must be indicated clearly.

***Note*. This excerpt includes remarks for test markers in order to raise its scoring validity**

I also depended on examples of excerpts from the actual test, i.e. excerpt 1 and rubrics of the test, to arrive at a rich and thick description and interpretations. Besides, I am, as a senior English teacher, familiar with the assessment processes, teachers' implementation of classroom assessment tools, and general education diploma exam marking.

***Excerpt 1*. Remarks for test markers**

Data Analysis

After reviewing the relevant literature, I reviewed and linked several official documents published by the MoE, analyzed the exam questions of the reading section and connected all the supporting evidences with Weir's socio-cognitive framework (2005). Table A1 displays the checklist that entails all the validities that make up the framework, the components under each domain and the subcategories whereas table B2 involves the documents being utilized and the way they served and supported the research argument.

Findings and Discussion

Test Takers Characteristics

Based on Weir's (2005) framework, three characteristics have to be accounted for: physical, psychological and experiential. Catering for physical characteristics relates to providing suitable accommodations for candidates with special needs and for individuals who may experience short-term sicknesses such as toothaches, earaches, etc. Item 33 in general diploma exam regulations and administration document (MoE, 2015) relates to specifying a room for candidates experiencing short-term health problems after obtaining the ministry's approval. Also, in the case of circumstances impeding the test takers from writing, the MoE allows having another person to write instead of the test taker provided that the writer is at a lower educational level than the test taker or an invigilator who does not teach the exam subject (MoE, 2015). However, the test taker admitted to a special room answers the same exam paper, without providing extra time nor interval breaks or reading assistance. Khalifa (2005) explicated that allowing extended time to students with disabilities has attested to improve students' test performance. Sitting for consecutively three hours to answer the whole test is undoubtedly exhausting and thus students' responses to the test questions may be impacted by tiredness they probably experience. Therefore, scores obtained can be questioned since they are probably affected by the length of the test and exhaustion of the test takers.

With respect to test takers' psychological characteristics, Weir (2005) postulated that "it seems unlikely that in the test event much can be done to cater for individual differences in these respects except to put the candidates at their ease as far as is possible" (p. 54). In other words, high-stake centralized tests like grade 12 exams have to be unified for all students across the country under very similar conditions

Table A1

A checklist of validity domains, components and subcategories

validity domain	components	subcategories
characteristics of test takers	physiological/ physical	• short-term sickness, i.e. toothache, cold • long-term disabilities in speaking, hearning, vision • age • gender
	psychological	• personality • memory • cognitive style • affective schemata • concentration • motivation • emotional state
	experiential	• education • exam preparedness and experience
context	task setting	• rubric • purpose • response format • known criteria • weighting • order of items • time constraints
	task demands	• discourse mode • channel of communication • length of text • nature of information in the text • content knowledge required • input/output (lexical, structural, functional)
	setting and administration	• physical conditions • uniformity of administration • security
theory-based	executive processes	• goal setting • monitoring • visual recognition • pattern synthesizer
	executive resources	• language knowledge (grammatical knowledge: lexical & syntax) • textual knowledge • functional (pragmatic) knowledge • sociolinguistic knowledge
	content knowledge	• internal (background knowledge) • external (task knowledge)
scoring	item analysis	analyzing item difficulty using statistical measurements
	internal consistency	measuring test homogeneity using statistical tests
	error of measurement	showing the differences between the observed score and the corresponding true score of proficiency
	marker reliability	• test type • number of raters • method of scoring
criterion-related	comparison with other tests/measurements	• scores from some other tests • candidate's self-assessment • teacher's ratings of the candidate
	comparison with an external benchmark	comparing test scores to external nationally accepted frameworks, i.e. council of Europe's common frame of reference
consequential	differential validity	• cultural background • background knowledge • cognitive characteristics • native language/ethnicity/age and gender
	washback	• impact on students • impact on teachers
	effect on society	• i.e. families

Note . I tabulated the validities and their sub-components from Weir (2005)

as such exams determine students' future tracks.

Checking the experiential characteristic of test takers, the MoE provides mock exams for all the subjects and grades. Students can also access the online test depository at the MoE portal. Additionally, most book stores in Oman sell booklets that include exam papers of all previous years along with answer guides. Though English teachers are urged to train their students with the types of exam questions all school year, students rarely sit for a full session that resembles actual exam setting using the mock exams provided by the ministry. We can understand that students are not sufficiently familiar with the exam experience particularly in the first semester since they do not sit for a mock exam session that reflects the actual exam experience.

Context Validity

In assessing the context validity of the reading questions, I firstly looked into task setting, secondly task demands and thirdly setting and test administration. Table 1A entails all the components and subcategories of this validity.

Task setting

The rubric for the three reading questions give clear instructions on what is expected from the students to do. The first reading question instructs students to "*Read the texts. Are the statements which follow each text* ***True*** *or* ***False****? For each item, shade in the bubble ( ) under the correct option*". The rubric is short with simple sentences, and gives clear indication of what the examiner is asking. No grammar or spelling mistakes are detected.

The second reading question consists of two related texts: a letter and a reply to it. Though it is divided into two separate texts, one rubric is provided for the two of them, " *Read Dr. Ali's letter to Mr. Smith and the reply. Then for each item, shade in the bubble ( ) next to the correct option*". It is not expected that having one rubric here would cause a confusion among the test takers because one kind of response, selecting the correct option, is required from the two texts' questions.

The case is different for the third reading question, where the examiners have provided separate rubrics for each task as different responses are required to answer the two tasks. For task one, the examinees are instructed to " *For each item, write a short answer (no more than FOUR WORDS).*" and for task two, the examinees are instructed to "*For each item, shade in the bubble ( ) next to the correct option.*".

As the rubrics of the three reading questions are unequivocal, it is expected that the students will execute the most appropriate strategies to answer the questions. Weir (2005) explicated that "having a clear purpose will facilitate *goal-setting* and *monitoring*" (p. 58). For reading comprehension questions, going through the questions before reading the text makes reading the text more purposeful, facilitates intelligent execution of time and most probably the test taker only needs to scan the text to respond to a specific question. However, further research needs to be done in order to uncover the strategies test takers employ while answering a test and this can probably be accomplished via the use of think aloud protocols, or post-tests questionnaires.

In totality, the test takers have to provide answers to twenty one items in the reading questions. In the first reading question, students should indicate whether the statements were true or false. This type of question can be ranked as easy and requires a low level of thinking since there is a 50% chance to get it right. The second question, students select the accurate answer from the three options given. It demands recognition too since the correct answer is provided and the test takers only need to shade it. Students only invest high level of thinking in the first task of the third reading question where they are asked to provide a short written answer which represents a high cognitive skill compared to recognition.

However, there is a heavy reliance on multiple choice questions (MCQs): ten MCQs with three options and seven with

two options to choose from. The level of guessing and getting some answers right based on mere guessing is high. Weir (2005) cautions using MCQs and true-false because: [t]he scores gained in MCQ tests, as in true-false tests, may be suspect because the candidate has guessed some or all of the answers. A candidate might get an item right by eliminating wrong answers- a different skill from being able to choose the right answer in the first place. If the answers are provided in this format, we never say whether a candidate would have got the item right without this assistance (p. 62).

Therefore, there is a possibility that the test response format is likely to affect the test performance and thus a variation of questions need to be considered that require various levels of cognitive and metacognitive strategies.

Clearly informing test candidates, as Weir (2005) pinpointed, about how they will be judged is of equal importance as having a clear idea of what they are expected to do in the test tasks. Samples of test papers of previous years are easily accessible through the Omani MoE portal (2020).Weir contends that "published information about how the tasks are scored, including criteria for correctness, steps used for scoring and how the item scores are combined into the test score, should be readily available" (p. 63). Excerpt1 shows remarks provided for test markers in the marking guide which is also publicly available online in the MoE portal. Therefore, test takers can have access to useful information regarding how to be judged in the case of committing grammatical mistakes, accuracy related issues and words limits.

Excerpt 1 goes here.

As there are twenty one question items, allocation of marks differs in the third reading question from the first two reading questions. Questions 1 to 13 worth one mark each, but items from 14 to 21 worth one-and-a-half mark each. Noticeably and very importantly, the test developers did not assign a specific time for each section. In this regard, Weir (2005) emphasized that: if different parts of the test are weighted differently, then the timing or marks to be awarded should reflect this and be evident to the test takers so that they can allocate their time accordingly in the goal-setting phase of processing. (pp. 63-64).

If the candidates are provided with the approximate time needed to complete each question, there is a probability of spending appropriate time for all exam portions, invest more time on the harder parts of the exam than the less demanding tasks and subsequently score better in the overall test.

Questions accompanying the texts in the reading part of the test require candidates to scan for correct answers. Thus, careful reading is not primarily required rather than random access to the text. Questions follow a serial sequence reflecting the matching answers as information appears in the text. Weir (2005) acknowledged this attribute and stated that "serial ordering of questions would progressively reduce the difficulty level of the exercise. If you [test takers] know the questions are in order, you would naturally not go back over what you had covered for the previous question" (p. 64). In short, test items in the reading task are in a justifiable order.

Task demands

The SAH (2016) stated that students are tested in unseen texts in their exams and thus one can argue that the texts used are authentic, a characteristic important when choosing an appropriate reading text for a test. Students are also familiar with different text types such as informative, narrative, argumentative or interactive texts, as indicated in both the SAH and the ELT curriculum framework (2012). Thus, the text types used in the achievement test, being validated in this paper, meet such a parameter. In short, the mode of the reading section is appropriate for the skills or strategies being tested.

The three reading questions vary in length. For instance, length of each text in the first reading question, ranges between 35 to 45 words. The second reading question involves two related interactive texts, with a length of 150 to 175 words. Students read an informative text with a

length of 425 and 475 words for the third reading question. Collectively, the texts in the three reading questions range from 970 words to 1185 words. Weir (2005) remarked that: the longer the text candidates are presented with, the greater the language knowledge that might be required to process it. If short texts are not making the demands on these resources that will occur in normal cognitive processing, theory-based validity is compromised. (p. 74)

Setting and test administration

Weir (2005) explicated that it is important to cater for the physical conditions of the test sites which concern for the "actual place, background noise, live or recorded materials, lighting, air-conditioning and power sources" (p. 83) particularly for listening tests. According to the MoE's document of test regulations and administration (2015), exam sites are elected by the educational general directorate for each governorate in the country. Test takers are not to exceed 300 candidates in each exam site, herein the schools, with about 15 students only in each classroom. Two invigilators are assigned for each classroom and they are instructed to not chat and to minimize their movements inside the class as to maintain a quiet atmosphere for the students.

For students in government schools, uniform instructions are provided for all the students, invigilators and test sites administrations across the country. For instance, in the test paper, students are given general instructions and guidelines on the front page only written in Arabic language. Considering the background of the test takers, all students are trusted to be able to read in Arabic in government schools. Students are expected to arrive at the test room 10 minutes prior to the beginning of the test time, which normally starts at eight and finishes at eleven in the morning. Arriving late to the test room later than ten minutes is not tolerated. Students have to show their identity card to the invigilators, but are not allowed to disclose their identities in the test paper by any means. Instead, they receive the exam booklets in sealed envelopes with a unique bar code for each student. Students are also prohibited from bringing anything to the test room including study notes, or any sort of devices or gadgets. Students are also reminded to write using blue or black biros only and to properly highlight the correct answers where appropriate.

The MoE's document of test regulations and administration (2015) includes several items that relates to the security of administering and invigilating the general education diploma across the country. Item 4 in the document concerns the selection of the invigilators who should be in a teaching-related profession and not to proctor in his/her original school. According to item 7, the test administrator has to arrange for a meeting with the invigilators prior to the start of the exams to communicate the regulations and procedures that highlight their responsibilities. Test invigilators have to declare any kinship they may have with the test takers according to item 8. According to item 14, carrying a mobile phone or any sort of computer gadgets are not allowed for all the test invigilators as well as others with administrative roles. The administrator is the only individual to receive the exam bundles on the day of the exam, and is instructed to open the exam bundles half an hour prior to the start time of the exam. All of these regulations, (see MoE, 2015), have been developed to ensure that test items are secured, and copying is unauthorized because "if tests are not secure, then some candidates may know the answers in advance and their processing will be of an entirely different nature" (Weir, 2005, p. 83).

Theory-based Validity

Apparently, the reading questions are of the expeditious nature at the local level and are primarily based on the skill of scanning since test takers are expected to locate specific pieces of information. Urquhart and Weir (1998) described scanning as a selective reading for the purpose of achieving specific reading goals, e.g. finding the number in a directory.

Short answer questions allow for a variety of skills to be tested such as inferencing,

recognition of a sequence, comparison and establishing main idea of a text. Seemingly, such types of questions require relating sentences in a text with other items which may be a distance away in the text. In short-answer questions, students are not provided with the answers as in the MCQs. Therefore, getting an answer right would undoubtedly reflect students' comprehension (Weir, 2005).

There are ten reading texts in the exam ranging in length from 35 to 475 words. Though it may look overwhelming, including such a large number of texts allows for a variety of topics to be covered and to reduce potential bias from using fewer topics (Weir, 2005).

Scoring Validity

In the test being validated, marks are assigned to all sections of the test, so candidates are aware of the mark allotted for every item. A comprehensive marking guide is provided for the test scorers coupled with marks for each single item. MCQs and true/false items are believed to eliminate markers' subjectivity. Whereas applicable, all possible answers, in the short questions items, are indicated for the test scorers, thus the mark scheme can be said to anticipate the kind of responses a test candidate is likely to make (see excerpt 1). Alderson (2000) noted that in a reading test "test designers should be open as possible in the range of different interpretations and understandings they accept" (p. 29) and this is evident in the remark given to the test scorers in excerpt 1 as to apply common sense to accept or reject answers not listed in the marking guide. Items in reading questions one and two are awarded with a mark for each response, while in the third reading question each item is awarded 1.5 marks. A candidate making a mistake in the third reading question will lose substantial marks which will eventually impact the overall score of the ELT. Though marks allocated for the first and second reading commensurate with the demands that the task makes on the candidate, the third reading question does not. A recommendation can be made to reduce the allotted marks for items in the third question.

The Omani MoE has adopted the electronic marking system for general diploma certificate examinations. Therefore, examiners need not to manually compute any marks for the candidate. They are only required to indicate whether the responses were correct or not. There are three markers for the exam which will likely minimize the human error while marking, raise inter-rater reliability and maintain the accuracy and consistency of the marking. However, it is a very long test (the full test is in 15 pages to answer in three hours) that students might lose their focus and concentration, and hence this may weaken their performance towards the end of the test.

Criterion-related Validity

Weir (2005) explained that judging a test to have a criterion-related validity refers to the presence of a relationship between test scores and "an external criterion which is believed to be a measure of the same ability" (p. 207). This can be achieved in two ways: 1. comparing the test with other tests/ measures, and 2. comparing the test with external benchmarks. Therefore, Weir contends that this type of evidence is hardly made available and that is why it is difficult to operationally put it into practice. Future research studies aiming to obtain a criterion-related validity in classroom settings can be conveniently done via comparing results of continuous assessment with the results of high stake end of semester national exams. Students' marks in English for grade twelve largely determine their future paths into tertiary education, though and due to confidentiality reasons no evidence could be accessed to make a comparison nor to draw a conclusion.

Consequential Validity

Test scores have huge impact at micro and macro levels. According to the Higher Education Admission Center in Oman (2019), more than 32% of grade twelve students in 2016/2017 who scored a D in English, a mark between 50 and 65, while they had a C as an overall grade, were not accepted into higher educational institutions. This is an example of how achievement in English language subject impact learners es-

pecially at critical educational stages. Test scores can affect various people and the society at large.

Firstly, at the level of the individual, test takers and teachers are both impacted with the testing process. Even though students receive their overall score in the English subject, they do not get qualitative feedback about their performance on the test paper. For any test to have a positive impact on the individual, feedback given to the test takers should be "relevant, complete and meaningful" (Bachman & Palmer, 1996, p. 32) and this can be achieved via providing test takers with feedback in the form of a score as well as a qualitative description of their performance, which gives meaning to the score. Tests can inform teaching and enhance students' learning; however, it can have a negative washback when it limits the teaching creativity and leads it to be test-oriented. Mohammed (2019) and Al-Lawati (2002) reported the impact of exams on the teaching instruction. For teachers, they were inclined to mainly focus on exam-oriented instruction towards the end of each semester, and for students, they admitted exerting their efforts to strategically prepare for the final exams. Also, centralized testing " implies a shift from student-centered to curriculum-centered instruction" (Runté, 1998, p. 167). At the level of society, many parents seek private tutoring for their children to get further preparations for the exams (Al-Issa & Al-Bulushi, 2012).

Conclusion

This paper has provided a detailed procedure of the validation process of the reading section of the GEDELT. Weir's (2005) socio-cognitive framework was employed and the research study based its argument on the foundation of some robust and pertinent documents mainly published by the MoE in Oman. Results of systematic evaluation of validity revealed that context-validity is not highly satisfied due to the test response format, absence of allotted time for each question and the exhaustion that the test takers may experience due to the considerable length of the test. Theory-based validity witnesses strengths by utilizing a large number of texts and a weakness due to overemphasis of the skill of scanning to locate specific information. Scoring-validity is considered to be high since types of task response, marking guides and electronic marking reduce markers' subjectivity and minimize human error. Similarly, the test in general and the reading section in particular show high level of impact on both micro and macro levels.

This study is only confined to the reading questions of the GEDELT in the Sultanate of Oman. It utilized Weir's (2005) socio-cognitive framework to validate a centralized local test. Though Weir's framework encompasses five validity components, only four were handled in this validation process because no evidence could be gathered to establish the test criterion-related validity due to confidentiality reasons. Also, this research paper relied on post-test evaluation and thus the gathered data formed the basis for data analysis and findings. It is advisable to conduct further studies to establish test validity and reliability using other quantitative and qualitative approaches.

References

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

Al-Ismaili, A. A. (2015). *Ensuring the context validity of English reading tests for academic purposes (EAP) in Oman* (Doctoral dissertation, University of Bradford, UK). Retrieved from https://bradscholars.brad.ac.uk/handle/10454/15710

Al-Issa, A. S., & Al-Bulushi, A. H. (2012). English language teaching reform in Sultanate of Oman: The case of theory and practice disparity. *Educational Research for Policy and Practice*, *11*(2), 141-176. https://doi.org/10.1007/s10671-011-9110-0

Al-Kharusi, H. (2011). Teachers' classroom assessment skills: Influence of gender, subject area, grade level, teaching experience and in-service assessment

training. *Journal of Turkish Science Education*, *8*(2), 39-48.

Al-Lawati, N. (2002). Washback effect of secondary certificate English examination on teaching and learning processes. (Unpublished master's thesis). Sultan Qaboos University, Oman.

Al-Mekhlafi, A., Al-Mamari, M., & Al-Barwani, T. (2019). The question of communicative language ability in EFL testing: The case of language testing in Oman. *Sumerianz Journal of Education, Linguistics and Literature*. 2(8), 51-61.

Angen, M. J. (2000). Evaluating interpretive inquiry: Reviewing the validity debate and opening the dialogue. *Qualitative Health Research*, *10*(3), 378-395. https://doi.org/10.1177/104973230010000308

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing*, *17*(1), 1-42. https://doi.org/10.1177/026553220017001010

Bachman, L. F., & Palmer, A. S. (1996). Language testing in practice: Designing and developing useful language tests. Oxford: Oxford University Press.

Bowen, G. A. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal*, *9*(2), 27- 40.

Carrell, P. L., Pharis, B. G., & Liberto, J. C. (1989). Metacognitive strategy training for ESL reading. *TESOL Quarterly*, *23*(4), 647-678. https://doi.org/10.2307/3587536

Creswell, J. W. (2007). Qualitative inquiry & research design: Choosing among five approaches. Thousand Oaks, CA: Sage.

Fulcher, G. (2010). *Practical language testing*. Uk, London: Routledge.

Higher Education Admission Center in Oman. (2019). *Number of general education diploma students who passed English language subject and got admitted into higher education in 2017/2018*. Oman: Ministry of Higher Education.

Hughes, A. (2003). *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Khalifa, H. (2005). Are test taker characteristics accounted for in Main Suite Reading papers? *Research Notes 21*, 7-10.

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

McNamara, T. (2000). *Language Testing*. Oxford: Oxford University Press.

Ministry of Education. (2012). *English language curriculum framework in Oman*. Oman: Ministry of Education Publication.

Ministry of Education. (2015). *Ministerial decree 588 regarding the regulations of managing general education diploma exams and what is at its level*. Oman: Ministry of Education Publications. Retrieved from https://home.moe.gov.om/file/karar w.pdf

Ministry of Education. (2016). *Student assessment handbook for English grades 11&12*. Oman: Directorate general of educational evaluation.

Ministry of Education. (2020). *Exams library*. Oman: Ministry of Education. Retrieved from http://zawity.moe.gov.om/images/library/file/Book6117328009.pdf

Ministry of Education, Oman & World Bank. (2012). *Education in Oman: The drive for quality*. Muscat, Sultanate of Oman: Ministry of Education. Retrieved from http://documents.worldbank.org/curated/en/246211468291645003/pdf/757190ESW0v10W0port0Summary-English.pdf

Mohammed, M. T. (2019). An exploration of why learning English for twelve years in Omani public schools is inad-

equate preparation for Omani students entering Higher Education. (Doctoral dissertation, The University of Liverpool, UK). Retrieved from https://livrepository.liverpool.ac.uk/3065894/1/H00025378_Nov2019.pdf

Plake, B. S., & Wise, L. L. (2014). What is the role and importance of the revised AERA, APA, NCME standards for educational and psychological testing? *Educational Measurement: Issues and Practice*, *33*(4), 4–12. doi: 10.1111/emip.12045

Rapley, T. (2007). Doing conversation, discourse and document analysis. London: Sage.

Runté, R. (1998). The impact of centralized examinations on teacher professionalism. *Canadian Journal of Education/Revue canadienne de l'education*, *23*(2), 166-181. https://doi.org/10.2307/1585978

Urquhart, A.H. and Weir, C.J. (1998). Reading in a second language: Process, product and practice. Harlow: Longman.

Weir, C. J. (1993). *Understanding and developing language tests*. London: Prentice-Hall

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Hampshire: Palgrave McMillan. https://doi.org/10.1057/9780230514577