Validation of an ESL writing test in a Malaysian secondary school context

Azlin Zainal

Faculty of Education, University of Cambridge, 184 Hills Road, Cambridge CB2 8PQ, United Kingdom

A R T I C L E   I N F O
Article history:
Received 18 October 2010
Received in revised form 16 August 2011
Accepted 19 August 2011
Available online 15 September 2011

Keywords:
Validation
Classroom-based assessment
ESL

A B S T R A C T
The present study was conducted with a twofold purpose. First, I aim to apply the socio-cognitive framework by Shaw and Weir (2007) in order to validate a summative writing test used in a Malaysian ESL secondary school context. Secondly, by applying the framework I also aim to illustrate practical ways in which teachers can gather validity evidence where this in turn would help them design and evaluate their tests in light of their teaching context and the purpose of assessment. In addition, teachers may be able to reflect on learners’ progress and areas where learners need to improve by looking at the interplay of tasks and learner’s response. Twenty exam scripts written by 16-year old ESL learners were rated based on a marking scheme to identify scoring validity. Finally, I will conclude that the validity of score interpretations has been established to a certain degree and the framework is practical for the purposes of the study.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

In the field of second language (henceforth, L2) writing assessment, the need for validation is seen as important especially given that, increasingly, direct assessment of writing is replacing traditional forms of indirect assessment. In some testing contexts, indirect assessment of writing through the use of gap filling, cloze test or multiple-choice test remains in use. However, there are contexts where such type of assessment has either been replaced with or supplemented by performance based assessment...
which involves judging learners’ performance in simulating a real world task. Among the factors that should be considered in assessing writing this way are the writing task, the written text, the scale applied, raters, writers’ characteristics apart from their writing ability and other contextual factors (Weigle, 2002). The importance of the social and cultural context in which the assessment takes place has also been stressed as the factors ‘may not be generalizable outside of that context’ (Weigle, 2002:60). This indicates the crucial need to understand the purpose of assessment in the context of use where recognizing the distinctions between summative or formative assessments would be a useful starting point to begin validity inquiry.

One of the purposes of classroom assessment is to measure learners’ achievement at the end of a year of schooling (summative). However, another purpose is to gauge students’ learning based on what has been taught and to provide appropriate feedback that would be useful to help learners to improve (formative). Viewing assessment as social practice, McNamara (2001) points out that validation demands made by researchers may not necessarily meet the needs of the teachers and learners and he suggests that ‘assessment specialists can help more adequately to theorize and conceptualize alternative, more facilitative functions of assessment in classrooms’ (McNamara, 2001:343). In other words, although validity of tests may be an important concern, another significant issue for assessment in pedagogical settings is how teachers and learners can use assessment as a way to systematically reflect on the outcomes of tests and use them as learning points in classrooms. Tomlinson (2005) goes to the extent of proposing that tests should also have ‘learning validity’ and suggests ways in which teachers and learners can learn throughout the process of preparation for tests, while sitting the test and from the feedback of the test. Although it is possible for validation procedures used by L2 testing researchers to be applied in classroom contexts, professional workload such as preparing materials and lessons, marking students’ work etc. may prevent teachers from engaging in extensive research. Furthermore, teachers may lack the proper tools or training on the sophisticated statistical procedures (McNamara, 2001). In view of this, I believe teachers would benefit from a framework that allows them to critically evaluate the effectiveness of the test they use in the light of their teaching context and at the same time use the data obtained from learners’ writing to inform decisions on feedback for learners. By applying the framework, teachers’ approach to test development and validation of their assessment would also be informed by an understanding of modern conceptions of validity. In this paper, I will first review the literature related to validation and L2 writing. Next I shall describe the method of gathering evidence for context, cognitive and scoring validity based on the socio-cognitive framework by Shaw and Weir (2007), where this is followed by a discussion of the findings.

2. Literature review

2.1. Unitary view of validity

Understanding the nature of validity should be our primary concern if we are to embark upon the exercise of validation. The current definition of validity is ‘the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of test’ (AERA, 1999:9). Prior to this view, however, the conceptualization of validity has not been without problems. No less than 16 types of validity have appeared in the psychometric literature since the 1930s (Cumming, 1996, citing Angoff, 1988). The most common types of validity commonly discussed at the time were content validity, concurrent validity and construct validity.

It has been argued that the previous disparate notions of validity are fragmented and lacking in comprehensiveness as they do not take into account the use of a test and its social consequences (Messick, 1989). To address this issue, Messick proposed a framework, referred to as the progressive matrix, where he argued that construct validity should be the unifying element for all attempts at validation. This unified view of validity has been endorsed by the AERA Standards (1999) where previously in the 1985 Standards it was not recognized. It is also important to understand how validity can be affected by other factors. Two of the threats that could affect validity include construct underrepresentation which ‘refers to the degree to which a test fails to capture important aspects of the construct’ and construct irrelevance referring to the extent that ‘the test scores are affected by processes that are extraneous to its intended construct’ (AERA, 1999:10). Traditionally, validity is also seen as distinct from reliability.
The more recent approach views reliability as evidence for validity. Naturally, we should be concerned with the scoring procedures and how reliable the raters are; however, it should also be mentioned that ‘reliability is a necessary but not sufficient condition for an ethical test’ (Hamp-Lyons, 1990:73).

2.2. A socio-cognitive framework

Validation of writing tasks and tests can be found in the literature (Read, 1991; Storch, 1993). The fact that we use scores to make decisions regarding individuals clearly highlights the need for ensuring that the scores obtained are valid. Messick’s framework has been used elsewhere (see, for example, Cumming & Berwick, 1996; Kunnan, 1998), and has proven to be an indispensable guideline for comprehensive validation of a test especially in large scale assessment contexts. For the current purposes, however, I utilised the framework by Shaw and Weir (2007). Adapted from Weir’s (2005) framework, it reflects a multifaceted view of validity. The choice of using this framework is not only because it is grounded in applied linguistics theory and research and applicable for classroom assessment. More importantly the framework appeals to a socio-cognitive approach in that ‘the abilities to be tested are demonstrated by the mental procession of the candidate (the cognitive dimension)’ and that ‘the use of language in performing tasks is viewed as a social rather than a linguistic phenomenon’ (Shaw & Weir, 2007:3).

The framework consists of two types of validation that is a priori or before-the-test event (context and cognitive validity) and also a posteriori which is after the test event—scoring validity, consequential validity and criterion related validity (see Appendix A). Shaw and Weir point out that ‘[t]he framework represents a unified approach to establishing the overall validity of the test’ (p. 3). In other words, the evidence generated for different types of validity are connected with each other; for example, context validity impacts on scoring validity as well as cognitive validity. The rest of the discussion in the literature review shall follow the framework (in particular, context, cognitive and scoring validity) and I will try to illustrate the link between the different types of validity and address the relevant issues related to both research in L2 testing and specifically L2 writing.

2.3. Context validity

The validity evidence we are interested in here refers to the types of tasks that we are setting the candidates. This is to ensure that the ability that we are trying to measure is not affected by the tasks themselves and learners can perform to the best of their abilities. It is also worth pointing out that ‘the characteristics of the test task are the only factors directly under our control as test developers’ (Bachman & Palmer, 1996:47). Knowing that we have a degree of control in task design, we should avoid testing irrelevant constructs. Nevertheless, it is noted that different candidates may respond to a task differently as they bring with them varying background knowledge and rely on various strategies to handle the task. Research into task-induced variability and task difficulty has been ongoing and at the moment remains inconclusive and there is still a need to understand how different tasks affect learners’ writing.

The literature also suggests that it is necessary to have guidelines on designing prompts for our tasks (Kroll & Reid, 1994). Various frameworks have also been proposed to describe task types highlighting the importance of understanding task characteristics. Bachman (1990) suggests that a framework of test facets is essential for validation of the tasks used in any kinds of tests, e.g. achievement, proficiency, etc. (p. 155). Task classifications specifically for writing have also been proposed by Weigle (2002), adapted from Purves, Soter, Takala, and Vähäpääsi (1984) and Hale et al. (1996). Shaw and Weir distinguish (2007) between ‘setting’ and ‘linguistic demands’ to generate evidence for context validity and this distinction is practical to follow (see Appendix A). The setting also includes test administration. This, however, will not be discussed in this paper as the focus is mainly on task setting and task demands.

There are several issues that have been considered contentious and these will be discussed briefly here. This includes the issue of whether or not learners should be given a choice of topics. It has been argued that by allowing learners to choose, they would be able to display their best writing ability. However, Weir (2005), in favour of the view of Jacobs et al. (1981), proposes that in writing
tests learners should be asked to write on the same topic. This is based on the premise that giving learners the choice of tasks would reduce the reliability on holistic scoring (Hughes, 1989). It has also been suggested that some tasks may be more difficult than others and this can also affect reliability. Furthermore, giving learners a choice of prompts is no guarantee that they will show their best writing ability, and learners may also lose valuable time in deciding the best prompt for them.

From the tester’s perspective, allowing learners a choice of topics could also pose as a potential source of construct irrelevant variance where this may reduce validity. An issue of concern is whether some groups of test takers would be disadvantaged when they are offered choice. Wainer and Thissen (1994) explored the issue of examinee choice in educational testing and asserted that test takers varied in their proficiency in choosing items well. Their review highlighted that in cases where the test items were not of equal difficulty, some groups of test takers (depending on gender and ethnicity) tended to choose more difficult items compared to others.

However, in a testing context where items are equated for difficulty, offering test takers a choice of topics may have a potential benefit. Jennings, Fox, Graves, and Shohamy (1999) investigated the effect of topic on the performance of ESL students with varying levels of English language proficiency in the context of the Canadian Academic English Language Assessment (CAEL). Controlling for task difficulty, they found no significant differences between the performance of the group of test takers who were given a choice of topic and the performance of the group who were not given a choice. Through the use of a questionnaire, the authors analysed the test takers’ perceptions of the factors that affected their test performance. The results showed that ‘choice’ was regarded by the test takers as the second most important factor after ‘time’. Although analysis of the choice patterns revealed that the low-proficiency test takers made choices that are different from the high-proficiency test takers, the authors suggest that offering the test takers a choice of topics did not place them at a disadvantage. Based on their analysis of the test takers’ response to the questionnaire, the authors conclude that ‘the more subtle and psychological effect of choice must not be overlooked’ (p. 451) and that providing test takers with a choice of topics may lessen their concern about topic effect. They recommend that the potential benefits of offering choice in L2 language testing is an area that deserves further investigation.

Another issue is response format. Bare prompts that do not provide much input to learners are not advisable as they may affect validity. Weir has suggested that in direct tests of writing, controlled writing is preferable as ‘this would facilitate marking and allow a more reliable comparison across candidates’ (2005:188). Other issues that are often discussed are related to the notions of audience and purpose in writing (Hamp-Lyons, 1990) with writing viewed as a social interaction (Nystrand, 1989, cited in Hyland, 2002). Considerations of purpose and audience are also found in Grabe and Kaplan’s (1996) taxonomy of writing skills. It is argued that knowing the audience and the purpose of writing will enable writers to write accordingly, thus producing better writing. Nevertheless, in examining the composing strategies of eight ESL learners, Raimie (1987) did not find any observable differences when audience and purpose were specified.

As mentioned previously, little is understood about task variables. It is, therefore, suggested that ‘increasing the number of samples of student’s work can help reduce the variation in performance that might occur from task to task’ (Weir, 2005:68). It can be suggested that for the context of summative assessment within secondary schools, the design of the tasks should primarily reflect the institutional learning goals. Whether or not we choose to give learners a choice of prompts or ask them to produce controlled or free writing, we should understand that we are primarily interested in assessing what learners have learnt. If we do decide to give a choice of prompts it has to be ensured that ‘the various tasks are as alike as possible in terms of cognitive demands, complexity of instructions, and so on, and that raters are trained to apply similar criteria on all tasks’ (Weigle, 2002:104). This emphasizes the interconnectedness of the validity concept—how the task design will impact on cognitive validity and also scoring validity; I will discuss these accordingly.

2.4. Cognitive validity

It has been acknowledged that there is a need to understand the construct of what we are measuring if we want to begin a validity inquiry. However, Bachman and Palmer (1996) suggest that construct can
be based on the syllabus (e.g. achievement tests) or, where there is none, on the theoretical framework of constructs (e.g. proficiency tests). Again this highly depends on the type of assessment.

Nevertheless, although the Malaysian syllabus does specify what learners should be able to achieve in terms of writing at the end of each year of their secondary schooling, we will remain true to the framework and review theories that attempt to describe the nature of writing. This has the purpose of linking the description of the processes that the L2 writer undergoes with what type of writing is required of them within the context of use. Models such as Bachman and Palmer’s (1996) and Grabe and Kaplan’s (1996) ‘provide taxonomies of the areas of linguistic and textual knowledge that are involved in language use’ and such models ‘along with the information about test takers, the testing context, and the purpose of a test, provide a useful starting point for defining the construct of interest’ (Weigle, 2002:79).

It should be noted, however, that currently there is no agreed definition on the exact construct of L2 writing. All that we can rely on is the descriptive processes (Grabe, 2001). Models of writing that are available in the literature are mainly based on first language (L1) writing, e.g. Hayes and Flower (1980), Bereiter and Scardamalia (1987). Based on these models we are made aware of the iterative nature of writing such as planning, drafting, and goal setting, which serves to show that writing is not a linear process.

In terms of L2 writing, the shortcoming of the models is that they do not address linguistic ability (Di Gennaro, 2006; Weigle, 2002). The models are based on the assumption that writers more or less have the same language ability, but this would be more appropriate in the L1 context. We have seen, however, from the applied linguistics perspective, models that attempt to explain the nature of language ability within the communicative context, taking into account linguistic competence (Bachman, 1990; Canale & Swain, 1980; Hymes, 1972). Grabe and Kaplan (1996) address the nature of writing ability specifically, by building on the previous models of communicative language competence and also by dividing language knowledge into three types: linguistic knowledge, discourse knowledge and sociolinguistic knowledge.

Weir (2005) proposes a theoretical model of L2 writing which incorporates the important elements of the above models. Based on this model, he argues that we can try to establish theory-based validity by ‘evaluating the activation of executive resources and executive processes prompted by the task’ (2005:110). Executive processes are said to be similar to the ‘strategic competence’ of Bachman’s (1990) model and they attempt to describe what writers do when they are engaged in writing. These processes should interact with their Executive resources which comprise linguistic and content knowledge.

Nevertheless, while this theoretical model is useful in describing the resources and processes that L2 writers are expected to draw upon in addressing a writing task, it is limited in that it does not include considerations of the different levels of processing that writers undergo. Shaw and Weir (2007) propose a cognitive processing framework based on the information processing models by Field (2004) and Kellogg (1994, 1996). This is an improvement to Weir’s (2005) theoretical model as the framework identifies the different phases and components of cognitive processes involved in writing. The parameters that are central to gathering evidence for cognitive validity are described by the authors (Fig. 1):

The framework by Shaw and Weir (2007) also includes consideration of how learners’ different types of proficiency can be catered for in the design of test tasks. They argue that the validity of a writing test can be established when candidates are engaged in all the processing components described ‘as appropriate to the level of proficiency being assessed’ (Shaw & Weir, 2007:42–43). The concepts of ‘knowledge telling’ and ‘knowledge transforming’ (Scardamalia & Bereiter, 1987) which provide an account for the difference between skilled writers and unskilled writers, can be applied in the description of task types. For example, it is suggested that less skilled writers do not engage in planning and revising strategies compared to skilled writers. One possible reason is that in producing a text, their attention is mainly directed towards trying to encode their thoughts in linguistic form, limiting them from developing cohesion and coherence of the text (Shaw & Weir, 2007). Although Scardamalia and Bereiter do not describe how writers progress from the stage of knowledge telling to knowledge transforming, Shaw and Weir argue that the concepts are useful in that they signify ‘a contrast between a linear writing process, presenting ideas as they occur, and one which entails organizing ideas in terms of their relationship to each other and to the goals of the text (2007:44). Tasks that involve lower levels of processing such as narrative writing entail a knowledge telling
process while tasks that demand writers producing arguments at higher levels involve knowledge transforming process.

As we can see, writing even in the L1 is considered complex, but it is more so when learners are writing in the L2. A good piece of writing will depend on learners’ ability to manipulate their knowledge resources effectively according to the requirements of the given task. Although little is known about learners’ cognitive processes in the context of timed assessment (Polio & Glew, 1996), it is acknowledged that learners would have an even more difficult task of monitoring the above processes within the time limit given and this should be taken into consideration when designing tasks.

The way we define the writing construct will also have an impact on what we should look for in learners' products and this also has a bearing on how we score learners' writing. What should be borne in mind is the focus of our assessment. McNamara (1996) distinguishes weak performance assessment from strong performance assessment. The former refers to the types of assessment that focus on language performance. A sample of language is elicited through the performance of tasks that resemble real world where learner's language proficiency is being judged. On the other hand, strong performance assessment will be judged on real world criteria. Learners are assessed based on their effectiveness in using language to achieve the aims of a communicative task according to real life situations (for example, to persuade, to reason, to negotiate). Understanding the continuum and where our tasks fit in, we should be able to apply a scoring procedure that represents the construct we are measuring.

2.5. Scoring validity

As mentioned previously, the construct of writing should also be reflected in the scoring procedures. In gathering evidence for scoring validity, we should first consider the types of scoring procedures or rubrics that are appropriate for a direct test of writing. Here we are also interested
in gathering evidence for the reliability of tests. We shall look at the issues surrounding L2 writing scoring procedures and raters separately.

2.5.1. Scoring procedures

Holistic scoring has been used continually since the 1970s in various contexts of L2 assessments. It involves assigning a mark or a percentage to an essay based on an overall impression of the essay guided by a scoring rubric. Among the advantages of holistic scoring is that it is said to be easy and practical to score. Holistic scoring rubric allows raters to concentrate on the strengths and not on weaknesses (White, 1984, cited in Weigle, 2002). According to Weigle, holistic rubric can be designed to allow raters to pay attention to aspects of writing that are viewed as essential according to the context, and ‘thus can provide important information about those aspects in an efficient manner’ (2002:112). Nevertheless, it has been proposed that holistic scoring, unlike analytic scoring, does not take into account learners' variability in terms of their L2 development (Hamp-Lyons, 1995). For example, some learners may have better control with organization of their ideas whereas some may have better control at grammar and the mechanics of the language.

Analytic scoring is different in that scores are obtained through evaluation of different aspects of writing (Brown & Bailey, 1984; Hamp-Lyons & Henning, 1991; Jacobs et al., 1981). Aspects of writing to be measured may include organization, content, vocabulary, mechanics and grammar. It is argued that analytic scoring gives a more complete picture of learners' writing ability. However, a disadvantage of analytic scoring is that it takes much longer to mark as raters have to go through a script several times. Furthermore, when all the scores given on separate aspects of writing are added up together, a single score would be given and all the information about the different aspects of writing is lost. In classroom-based assessment, analytic scoring would help learners for diagnostic purposes or as feedback on which aspect of writing they should improve on. However, there is also the issue of designing the appropriate scales and the question of transferability of the existing scales to a different context (Hamp-Lyons & Henning, 1991).

2.5.2. Raters

It has been suggested that using holistic scores may be problematic for raters, especially in trying to balance all aspects of the given criteria. If a learner’s script has glaringly obvious errors in terms of grammar, raters may overlook other important aspects such as vocabulary or syntactic complexity. This also brings us to the question of inter- and intra-rater reliability. For the scores to be reliable, raters will have to be consistent with their interpretation of scoring criteria and assign marks accordingly. This consistency also has to be maintained between raters to avoid adverse effects on reliability. It has been shown, however, that with rater training and standardization, reliability can be achieved (Carlson et al., 1985; Homburg, 1984; Reid, 1993; Upshur & Turner, 1995). Studies on raters' rating processes using verbal protocol analysis (e.g. Cumming, 1990; Huot, 1993; Weigle, 1994) suggest that ‘experienced raters utilize strategies in approaching the scoring criteria and also in monitoring themselves in the process of rating’ (Sakyi, 2000:131).

Having looked at issues surrounding scoring validity, the questions that one needs to ask here in terms of rating scale are whether the existing rating scales function well for the intended purposes and context, and as for raters, whether they have been trained to apply the scoring procedures accordingly.

2.6. Research questions

I aimed to gather evidence in the following areas: context validity, cognitive validity and scoring validity. These are important to establish before applying external validation procedures to establish consequential validity (Weir, 2005). In my validation, I attempt to answer the questions that Shaw and Weir (2007) argue are necessary for test developers or users to address. They are as follows:

1) Are the contextual characteristics of the test tasks appropriate and fair to the candidates?
2) Are the cognitive processes required to complete the tasks appropriate?
3) How far can we depend on the scores of the test for the purpose we want them to serve?
3. Method

3.1. An action research approach to validation

I adopted an action research framework for the purpose of this research. As pointed out by McCutcheon and Jung (1990) ‘[a]ction research is characterized as systematic inquiry that is collective, collaborative, self-reflective, critical, and undertaken by the participants of the inquiry’ and that one of its goals is ‘the understanding of practice’ (p. 148). The outcomes of this research are not intended to only benefit the research participants but also language teachers in general. In applying Shaw and Weir’s framework, this research aims to address issues related to classroom practice in terms of designing and validating language tests based on teachers’ context of teaching and the purpose of testing.

An element of collaboration between the researcher and the participants characterized the approach to this research. The data was collected with the assistance of the main participant of this study—the class teacher. The class teacher, a Malay and a female, received her BEd in Teaching English as a Second Language from Warwick University, United Kingdom and had four years of teaching experience. She was a former colleague as I had previously taught at the school where the data for this study was obtained. The learners who took the test were studying at MARA Junior Science College (MJSC) Muar, a boarding school. Majlis Amanah Rakyat (commonly abbreviated as MARA) which translates as the People’s Trust Council is responsible for the planning and the monitoring of educational activities in MJSCs. The school, which was opened in 1981, is one of 49 MARA Junior Science Colleges in Malaysia.

3.2. Curriculum specifications for Form 4

The status of English in Malaysia is widely accepted as that of a second language. The Form 4 syllabus specifies the context of language use into language for interpersonal, informational and aesthetic use rather than separating the four skills of speaking, listening, reading and writing. The themes and topics through which the uses of the language should be taught are also listed. Generally the themes include the broad headings of ‘People’, ‘Environment’, ‘Social Issues’, ‘Values’ ‘Health’ and ‘Science’ and ‘Technology’. It is noted that learners are not expected to have ‘a comprehensive academic knowledge of these topics’ (MOE, 2003:8).

3.3. Test takers’ characteristics

It is important to recognize the test taker characteristics in terms of physical/physiological, psychological, and experiential (O’Sullivan, 2000, cited in Weir, 2005) as these will have an impact on both task design (context validity) and learners’ cognitive processes (cognitive validity). The process of student selection upon entry to MARA Junior Science Colleges is based on their Pendidikan Menengah Rendah (which translates as Lower Secondary Assessment and commonly abbreviated as PMR) results, a national exam taken in Form 3 at the age of 15. Students also come from all over the country. It is anticipated that the backgrounds and previous education of the learners who took the test vary somewhat which could possibly explain their mixed levels of English language proficiency.

Being proficient in English is considered an advantage as it is seen as a means of advancement in further education. Towards the end of their schooling, students have to sit for another national examination called Sijil Pelajaran Malaysia (SPM) or the Malaysian Certificate of Education. This examination is considered important as entrance to higher education is contingent upon the results. A pass in the SPM English 1119 paper is also one of the basic requirements to enter higher level education.

3.4. Summative assessment of Form 4 students

The summative assessment of Form 4 students is based on the SPM English 1119 format. This is different from the format of PMR for which they had sat when they were in Form 3. However, this group of learners should be familiar with the SPM format by now and how papers are marked as they had sat for the mid-term exam paper which had similar format. Learners would be aware of the
rubrics, the demands of the different types of tasks and the scoring procedures. This is important to allow learners to have equal chances of success.

The questions were designed based on the questions pooled from the English Unit of each MARA school. Every year, the MARA English Curriculum Unit will require each school to send a set of exam questions for them to select and use for the standardized summative exams. The questions and marking scheme were developed by the MARA English Curriculum Unit to have the same form as SPM English 1119. The paper is standardized and administered in all MARA schools.

3.5. Test description

The grade for the English 1119 paper is obtained based on performance on two papers; Paper 1 and Paper 2. I shall not describe Paper 2 in detail here as it mainly consists of multiple choice and subjective questions. However, it is necessary to describe what Paper 1 entails as learners’ writing is assessed based on their performance in this paper.

This paper consists of two sections. In Section A (Directed writing), learners are required to produce an essay with a specific format (i.e., articles, debates, letters, reports, etc.). Learners are expected to produce the correct format and also use any of the points or phrases given in the prompt. In Section B (continuous writing) learners are given a choice of bare prompts which would require them to produce different modes of discourse (i.e., argumentative, descriptive, discursive, expository and narrative).

3.6. Instrumentation

I used the framework by Shaw and Weir (2007) to evaluate the tasks set for the learners. From the perspective of the test user, I tried to establish the context and cognitive validity by comparing learners’ responses with the questions set. In gathering evidence for scoring validity, exam scripts of twenty Form 4 (16-year olds) students were rated by the class teacher and the researcher. The exam was part of the end of the year school exam. With the exception of two students (a Chinese and an Indian), the rest are of Malay ethnicity and their first language is Bahasa Malaysia.

As mentioned previously standardization of marking among teachers can be achieved (Weigle, 1994). Training is often given to new teachers during an induction course organized by MARA to mark scripts using the SPM English 1119 marking method. At the school level, the teachers from the English unit would gather and discuss the marking scheme and also look at samples of learners’ scripts to align marks according to scoring descriptors. As I had taught there myself, I would say that I am familiar with the marking method. However, unlike before, I have had no opportunity to discuss with my colleagues to align my scores. Therefore, I looked at a few samples of writings that represent a different mark range as a baseline before marking them and the rest of the paper again. Using Microsoft Excel, Pearson’s correlation coefficient was calculated based on the sets of scores (see the discussion on scoring validity for the correlation coefficients).

4. Results and discussion

4.1. Context and cognitive validity

I will not separate my discussion of context and cognitive validity in this section. As noted earlier, the types of task we set the test takers will have an impact on how they perform. In other words, we can see how the task characteristics interact with learners’ cognitive processes by looking at learners’ compositions. However, caution has to be taken here as I can only attempt to describe rather than provide explanatory arguments as to how learners react to the different prompts given that I did not use methods that actually tap into learners’ thinking, e.g., a think aloud protocol or questionnaire. Therefore, I will describe the content of the paper and identify how learners’ writings are affected by the prompts where this will also contribute towards further understanding of the effects of task variables. In generating evidence for both context and cognitive validity, the discussion will be based on the contextual and cognitive parameters set out by Shaw and Weir (2007) (see Appendix A). In
addition, some areas of scoring procedures will be mentioned to illustrate, where relevant, how the construct is reflected in the scoring procedures.

Looking at the content, I would say that the rubrics of the paper are clear and accessible in terms of the types of question, time, and marks to be awarded for each section. Learners are asked to produce two samples of writing and this can be argued increases the validity. I shall now discuss each section separately.

4.1.1. Section A

It is clearly stated at the top of the paper that they are required to write a piece of Directed Writing. Learners were familiar with the term ‘Directed Writing’ as they had been exposed to how to answer this type of task. The time given is 45 minutes and the full marks awarded are 35 (see Appendix B).

The purpose of writing is clear. Learners are given the role of the President of the Student Body of their school and asked to write a report to the school Principal about the deteriorating conditions of the school and recommendations on how to improve the situation. Learners are made aware of the audience who will be reading this (the Principal) as invoked by the task.

This is a controlled writing task. In terms of macro planning and organisation, limited planning is required by this task as it provides scaffolding in terms of input and structure. Learners are given six points (e.g., broken school equipment, leaking roofs, peeling paintwork, etc.) to elaborate on. They are also asked to give a title to the report, add two recommendations and provide relevant details. Since the task requires learners to write a report, learners should be able to activate their cognitive processes and recognize the format or genre of a report. They should include a title as part of the report format and are also expected to be able to use the given points. The word ‘all’ is highlighted in bold to emphasize the importance of each point. This is because failure to use a point will result in a mark being penalized.

In terms of linguistic demands, the discourse mode can be said to be appropriate. ‘Writing brief reports’ is also a curriculum specification (MOE, 2003:18) and learners would have been taught how to write a report in class. Although the number of words is not specified here, there is a degree of flexibility as the length is determined by how much learners could expand on the points. Learners can also use the recommended time for them to complete the task as an indicator of how much they should write. Furthermore, since there are not many points for learners to elaborate on, learners of both high and low levels of proficiency should be able to complete the task within the specified time limit. The topic is also relevant for learners at secondary school and the points include vocabulary related to the school environment.

The functions in the test input are appropriate for the level of the learners and for producing the output. The instructions and points given should enable learners to ‘generate ideas within the specific framework provided’ (MECU, 2006:2) and to produce a report as the output of the text. However, the lexical items in the input could again affect the overall outcome. In one sample, a learner misinterpreted the word ‘deteriorating’, thinking that it refers to a programme and has positive connotations. This affects the learner’s elaboration of the given points. Instead of a report about the worsening conditions, this learner actually suggested that the “programme” should be continued.

It is also noted that in terms of the format, some learners used a format which is different from that of a report. For example, in one sample the learner only wrote the title but not another essential element of a report which is ‘Written by, signature’. Perhaps, the fact that the question only required the learner to write the title in terms of format, prompted the learner to do so where this does not really show that this learner has mastered the format of the report. Nevertheless, only one case is evident while the rest of the learners did put their signature even though a mark is not awarded for that. The title should also be appropriate because a title that did not reflect the report as whole was not awarded a mark. An example of such title is “Recommendations of school”.

In terms of translation, encoding their thoughts into linguistic form involves the learners using ‘clear and accurate standard English in the response, using a style and tone appropriate to the task’ (MECU, 2006:4). The use of formal language is required here and scripts that fail to comply with this may not receive high marks. Learners’ ability to convince the reader that it could be a report is an example of a strong performance assessment (McNamara, 1996).
The learners' knowledge of the criteria of how the task is scored might affect their monitoring and revising strategies. One of the assessment objectives is to ‘use the information given to display an understanding of the task’ (MECU, 2006:2). The lexical and grammatical items used would perhaps serve to differentiate between the more proficient and less proficient students. However, although most of the items were well understood by the learners based, an often-misunderstood phrase is “unkempt compound”. Learners who were not sure of the meaning of the phrase were either unable to elaborate on the point or misused it altogether, e.g. “unkempt compound should be repair”, “ask the rubbish lorry to come to take the unkempt compound” and “the toilets are full of unkempt compounds”. The last example, however, unexpectedly came from quite a proficient learner mistaking the word compound for substance. This serves to highlight the crucial issue of selecting lexical items that are appropriate to the linguistic levels of the learners.

4.1.2. Section B

In contrast with the previous section, Section B is a free writing task. Learners are given a choice of prompts. Again, in terms of rubrics, it is clear what learners are expected to do. Learners should also be aware of the heavier weight on this section, with 50 marks clearly stated at the top of the test paper. Having done this type of writing, learners should have had the experience to plan their writing and finish it within the suggested time limit which is one hour. The length of the composition is expected to be not less than 350 words. The topics include the following:

a) My favourite person
b) Travelling broadens the mind
c) Write a story beginning with: “Suddenly the music played….”
d) What would I do to help save the environment
e) Dreams

The task demands are higher on learners' cognitive processing compared to the previous task. The prompts do not provide much input and learners would need to be able to access and activate their knowledge resources based on the prompts. It is suggested that ‘[i]n constructing tests it is important to include texts and activities that mirror as closely as possible those that students have been exposed to and/or are likely to meet in their future target situations’ (Weir, 2005:61). The prompts reflect the genre of writing learners are familiar with and should be able to perform according to the requirement of each task. For example, the prompts require learners to write (a) a descriptive essay, (b) an expository essay (c) a narrative essay and (d) a discursive essay. Task (e) however, is open to learners’ interpretation. Ability to write different types of text is also in the curriculum specifications and learners would have been taught this in class. The lexical and grammatical items in the input also appear unproblematic. By giving a choice of topics based on different genres of writing, learners of different levels of proficiency may be able to choose a topic that best displays their writing ability. Writers of lower levels of proficiency may find the tasks of writing descriptive essay and narrative essay (which involve ‘knowledge telling’) easier compared to writing expository or discursive essay (which entail ‘knowledge transforming’).

Unlike in the previous task, the purpose here is not stated. Nevertheless as suggested by Grabe and Kaplan (1996), the purpose of writing should be addressed at two levels. Generally, the purpose of writing in this task ‘is related to the concept of genre’ where the “purpose” influences the text structure itself, selecting appropriate genre options’ (1996:210). In other words, the purpose is embedded in the text that should be generated. In terms of output, learners are expected to produce texts that have the features of the above genres. For example, they should recognize that the structure of a narrative with the purpose of telling a story will not be similar to the structure where the purpose of writing is to discuss how travelling broadens the mind. In marking learners’ scripts, teachers should pay attention to learner’s writing characteristics—whether or not they reflect the genre of the task.

Another level of purpose described by Grabe and Kaplan relates to functional issues such as to apologize, to report etc. is not required by the task. Audience is also not specified by the task. However, for the learners, an audience consideration (Grabe & Kaplan, 1996) may be the teacher who would be marking their essays and who is also known to the learners. Given the scores are to be interpreted for
internal purposes, the task can be considered as ‘interactionally authentic’, with the learners addressing the teacher. However, it is noted that the scores might not generalise well to real world tasks with other audiences. Although this makes the task interactionally less authentic, it does not necessarily contribute to the invalidity of score interpretation. What should be borne in mind are the assessment objectives, or the construct that we are measuring. Through this task, learners are expected to ‘produce a piece of continuous prose in accurate standard English’ and ‘respond relevantly and creatively to a task chosen from a number of alternatives’ and particular emphasis is placed on several aspects of writing, i.e. sentence structure, vocabulary, punctuation, spelling, paragraphs etc. (MECU, 2006:5). Based on the holistic descriptors, raters should pay attention to grammatical and discourse aspects of writing. Functional or sociolinguistic knowledge (related to audience and purpose) are not being tested here. This is also indicative of performance in the weak sense (McNamara, 1996).

In terms of response format, open ended essays are considered problematic as ‘an ability to write on such open ended topics may depend on the candidate’s background or cultural knowledge, imagination or creativity’ (Weir, 2005:164). Learners are expected to engage in their knowledge resources and access their internal content knowledge to generate, organize, and translate ideas into a piece of writing that is relevant to the topic given. However, the topic selection was not random for they were based on themes listed in the curriculum specifications above. Learners are not judged based on their background knowledge. The content is functional only to allow learners to display their grammatical and discourse knowledge where this will be judged on the accuracy, relevance and creativity of the writing. A problem perhaps for raters includes identifying what can be considered as creative as it is not clear in the marking scheme what aspect of creativity should be looked at. With creativity one could imagine ideas or language but since we are not evaluating learners’ ideas, we are left with looking at creative use of language. This is perhaps difficult to establish in essays of expository or argumentative nature.

The task does not include direct reference on how learners should organize their responses. Unlike in Section A where the structure of the task is outlined, this task demands learners to structure their writing based on their knowledge on how to build cohesion and coherence in writing. The task also places greater demands on the learners’ processes of monitoring and revising. The learners need to monitor their writing in terms of mechanical accuracy such as grammar, syntax and be able to revise any errors accordingly. At a higher level, learners should address the goals of the task and that their writing should reflect the topic chosen. In one sample, a learner failed to address the goal of the task by misinterpreting the title ‘My favourite person’ to mean ‘My ideal guy’ and wrote an essay that is irrelevant to the topic. The marking scheme does state what should be done with such ‘abnormal scripts’ (MECU, 2006:7) and raters should be aware of how to score such scripts.

4.2. Scoring validity

I scored the essays based on the provided scoring rubric. The marks for the first task were awarded based on three categories: format (1 mark), points (14 marks) and language (20 marks). This is a form of analytic scoring. Nevertheless, the language aspect was not looked at in sections but instead awarded based on impression marking according to the descriptors. For the second task, the essays were scored based on a holistic scale (50 marks).

4.2.1. Comparison between Section A and B

It has been suggested that the use of controlled writing would allow more control on the different variables (Weir, 2005). This is reflected in the results of the Pearson correlation test. We can see that for Section A, there is a high correlation between the teacher’s scores and my scores, \( r = 0.98 \), while Section B has a slightly lower correlation between the teacher’s scores and my scores of 0.95. The significance of the results cannot be established given the limited samples. Nevertheless, overall the results suggest a high inter-rater reliability for both tasks thus showing that reliability can be achieved with the use of marking schemes, training in how to use them and standardization. Having gone through the marking exercise, several issues can also be highlighted (Figs. 2 and 3).

Marking Section A is more time consuming as raters are expected to check that the format is right, read to see whether learners have used the given content points, ensure that elaboration is
relevant to the content point given and read again to score the language. For the content point, marks can be awarded as long as learners mention the content points in a grammatically correct way and this is basically unproblematic. This is also an advantage to the least proficient learners as they are still awarded a point although they may not understand the meaning of the given phrase. What is potentially difficult is the elaboration, as raters have to decide whether it merits a mark. Another is their recommendations for other areas of improvements. Some learners misinterpreted this with suggestions like “collect donations,” “have a canteen day to raise funds”, etc. Although these are reasonable suggestions, the recommendations required by the task are specific to how else the school should be improved and not how money could be raised to help to improve the school. The prompt only states ‘add two recommendations’ as it assumes the learners will be able to link it back to the question and understand what is required. This is a disadvantage to learners who misinterpreted it.

For Section B, although the correlation is high, the fact that it is lower than Section A may warrant explanation. Response format or the rating scale may contribute to this although it is beyond the scope of this essay to establish this as other complex statistical measurements will be needed. Nonetheless, as Weigle points out, holistic scale is ‘not always easy to interpret’ as ‘raters do not necessarily use the same criteria to arrive at the same scores’ (2002:114). One of the questions that arises is when learners misused the tense. In describing a person, we would naturally expect learners to use a present tense.
However, in one sample describing Princess Diana as a favourite person, the learner did not use past tense to describe her and did not conclude the essay. The class teacher commented that past tense should be used since the Princess had passed away, and gave the essay 25 marks. Nevertheless, despite the learner’s failure to use past tense, the learner’s use of vocabulary, could perhaps have entitled her to a slightly higher mark. Here is an example where having an analytic score would help to give an idea of learners’ strengths and weaknesses. This is important as the mark given is supposed to represent the writing as a whole. Therefore, feedback on how the learner did well in other areas is possibly a better approach in this case and perhaps discussing it among raters and agreeing on what should be done with such script would be helpful. However, as long as raters are aware of these issues, there should be no reason why an agreement could not be reached on how to rate such script where this would further contribute towards inter-rater reliability.

5. Conclusion

Although the test did not meet all of the requirements of Shaw and Weir’s (2007) framework, the context, cognitive and scoring validity all have been achieved to a certain degree to support the interpretations of scores for the intended purpose. As Messick pointed out, ‘validity is a matter of degree, not all or none’ (1989:33). By examining the contextual characteristics of the test tasks in terms of the task setting and linguistic demands, it can be suggested that they are appropriate for the levels of the learners in this study. With regard to cognitive validity, the characteristics of the test tasks appear to activate the cognitive processes which are appropriate to the proficiency levels of the learners. Descriptions of learners’ writing also shed light on how the tasks affected learners’ responses. Evidence for scoring validity was achieved based on the high correlations between the teachers’ scores and the researcher’s scores established through Pearson correlation test.

The controlled writing task and the analytic scale met most of the demands of the framework but the free writing task and the holistic scale may not be ideal according to the framework. However, it should be noted that this is part of a summative test at a school level and not a high stakes proficiency test used for gate-keeping purposes where the design of the latter would normally be based on theoretical constructs. Therefore, it can be argued that if the tasks reflect the institutional learning goals and learners have been exposed to the types of tasks required of them, and both raters and learners are aware of how essays are judged, there should be no reason why validity of interpretation of scores cannot be achieved. Nevertheless, it should be cautioned that it would be less valid to interpret the scores as indicators of how well the learners are able to perform in real world tasks or with other audiences given their familiarity with the task types and their knowledge of what is expected of them.

I can also suggest that the framework is practical for classroom-based assessment validation. Once teachers are familiar with the issues related to the framework, they would be more aware of what should be paid attention to. Furthermore, by examining learners’ responses to the task, teachers will not only be able to measure each learner’s achievement but could also highlight areas that require more work and use them as teaching points in future lessons. Although summative tests primarily function as measures of achievement, there should be no reason why learners’ products should not provide information about learners’ progress as formative assessments would.

This study, however, has primary implications for the test user in pedagogical settings. As for the test developer (e.g. MARA Curriculum Unit), a validation research such as this could also be taken further to include larger samples of writing to establish criterion related validity or investigate the effects of task variables on learners’ writing. It is hoped that further investigations into a potentially rich area such as classroom-based assessment could promote stronger links between the fields of language teaching and testing.

Acknowledgements

I thank Alastair Pollit for supervising me in this study. I would also like to thank Hakimi Mohd Yusof and Meg Zellers for their assistance in proofreading earlier drafts of this paper and the three anonymous reviewers for their constructive and helpful comments on my manuscript.
Appendix A.

A socio-cognitive framework for validating writing tests (Shaw & Weir, 2007:4).

<table>
<thead>
<tr>
<th>Context validity</th>
<th>Demands: Task</th>
<th>Cognitive Validity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Setting: Task</td>
<td>Linguistic (Task input and output)</td>
<td>Cognitive processes</td>
</tr>
<tr>
<td>Response Format</td>
<td>Lexical resources, Structural resources, Discourse mode, Functional resources, Content knowledge</td>
<td>Macro-planning, Organisation, Micro planning, Translation, Monitoring, Revising</td>
</tr>
<tr>
<td>Purpose</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knowledge of criteria</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weighting</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Text Length</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time constraints</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Writer-reader relationship</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Setting: Administration

Physical Conditions

Uniformity of administration

Security

<table>
<thead>
<tr>
<th>Test taker characteristics</th>
<th>Scoring validity</th>
<th>SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Physical/Physiological</td>
<td>Rating</td>
<td></td>
</tr>
<tr>
<td>- Psychological</td>
<td>Criteria/rating scale</td>
<td></td>
</tr>
<tr>
<td>- Experiential</td>
<td>Rater characteristics</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rating process</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rating conditions</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rater training</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Post-exam adjustment</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Grading and awarding</td>
<td></td>
</tr>
</tbody>
</table>

Appendix B. Section A: Directed Writing

[35 marks]

You are advised to spend about 45 minutes on this section.

**Question 1**

You are the President of the Student Body of your school. You noticed the deteriorating conditions of the school. Write a report to your Principal with recommendations to improve the situation.

In your report you should include all the following:

- Broken school equipment
- Leaking roofs
- Peeling paintworks
- Cracked floor
- Clogged drains
- Unkempt compound
When writing the report, you should remember to:

- Give your report a title.
- Use all the points given.
- Add two recommendations.
- Provide relevant details.
- Write in paragraphs.

References


**Azlin Zainal** is a PhD candidate at the Faculty of Education, University of Cambridge, UK. She is also a researcher at the Faculty of Languages and Linguistics, University of Malaya, Malaysia. Her research interests include ICT in language education, teacher education and second language writing.