العوامل المؤثرة في الاختبار المباشر لمتعلمي اللغة الثانية القدرة على الكتابة

أنور سعد الجدعاني

أستاذ مساعد في علم اللغة التطبيقي - قسم المواد العامة

كلية الآداب والعلوم الإنسانية

مستخلص. يسلط المقال الضوء على بعض العوامل المؤثرة في اختبار مهارة الكتابة المباشرة لدى متعلمي اللغة الثانية. يعرض أهمية أربعة عوامل يمكن أن يكون لها تأثير في الاختبار المباشر لقدرات الطلاب على الكتابة. العوامل هي الوضوح، و الصلاحية، و الموثوقية، والمتعة والاهتمام. الوضوح ويقصد به أن يعرف المتعلم المطلوب منه بشكل سهل وسريع وواضح. أما الصلاحية فتعني أن يقيس الاختبار المهارات المطلوبة بدقة، والصلاحية لها أربعة أنواع: صلاحية الشكل، وصلاحية المحتوى، وصلاحية المعيار، وصلاحية الإنشاء. العامل الثالث هو الموثوقية، وهي أن تكون النتيجة ثابتة من اختبار إلى اختبار. العامل الأخير هو المتعة والاهتمام وذلك بأن يكون موضوع الاختبار من اهتمامات المعلم والمتعلم. اختتم المقال بأهمية رجوع مصممي الاختبار إلى تقييم جميع العوامل المتنوعة خلال مرحلة بناء الاختبار واختيار العوامل التي ستؤدي إلى اختبار عادل وصحيح يكون واضحاً وموثوقاً به ومثيرة للاهتمام لجميع الأطراف المعنية.

to weigh up all the different factors during the test construction phase and choose those that will result in a fair and valid test that is clear, reliable and interesting for all parties concerned.

References

- Anastasi, A. (1986). *Evolving concepts of test validation*. Available from www.annualreviews.org. (Accessed 17 April 2011).
- Cronbach, L. J., and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302.
- Douglas, D. (2000). *Assessing language for specific purposes: theory and practice*. Cambridge: Cambridge University Press.
- Fulcher, G. and Davidson, F. (2007). *Language testing and assessment: an advanced resource book*. New York, Abingdon: Routledge.
- Hamp-Lyons, L. (1989). Raters respond to rhetoric in writing. In H.W. Dechert and M. Raupach (Eds.), *Interlingual processes* (pp. 229–244). Tübingen: Gunter Narr Verlag.
- Hamp-Lyons, L. and Kroll, B. (1996). Issues in ESL writing assessment: an overview. *College ESL*, 6(1), 52–72.
- Hamp-Lyons, L. (2000). Social, professional and individual responsibility in language testing. *System*, 28(4), 579–591.
- Hamp-Lyons, L. (2003). Writing teachers as assessors of writing. In B. Kroll (Ed.), *Exploring the dynamics of second language writing* (pp. 162–189). Cambridge: Cambridge University Press.
- Hinkel, E. (2002). *Second language writers' text: Linguistic and rhetorical features*. Mahwah: Erlbaum.
- Hinkel, E. (2009). The effects of essay topics on modal verb uses in L1 and L2 academic writing. *Journal of Pragmatics*, 41, 667–683.
- Huot, B. (1996). Toward a new theory of writing assessment. *College Composition and Communication*, 47, 549–566.
- Hughes, A. (1989). *Testing for Language Teachers*, 1st ed. Cambridge: Cambridge University Press.
- Hughes, A. (2003). *Testing for Language Teachers*, 2nd ed. Cambridge: Cambridge University Press.
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher* 23(2), 13–23.
- Messick, S. (1996). Validity and washback in language testing. *Language Testing* 13, 241–256.
- Miles, M. and Huberman, A. (1994). *Qualitative data analysis: an expanded sourcebook*, 2nd ed. Thousand Oaks: Sage Publications.
- Nevo, B. (1985). Face validity revisited. *Journal of Educational Measurement*, 32(4), 287–93.
- Wall, D. (1998). Impact and washback in language testing. In C. Clapham (Ed.), *The Kluwer encyclopedia of language in education*, Vol. 7: testing and assessment (pp. 334–343). Dordrecht: Kluwer.
- Wall, D. and Alderson, J.C. (1993). Examining washback: the Sri Lankan impact study. *Language Testing*, 10(1): 41–69.
- Weigle, S.C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.
- Wiseman, S. (1949). The marking of English compositions in grammar school selection. *British Journal of Educational Psychology*, 19 (3), 200–209.

that specific person's performances. This may be especially true when dealing with L2 students as even more variables come into play. It is further complicated by the fact that writing assignments are judged by human beings and these human beings are also likely to vary from day to day and from subject to subject; they are likely to have preferences for certain kinds of ideas or structures, or dislike some choices of words or arguments (Hamp-Lyons, 2003).

How then can we ensure at least some degree of reliability in direct testing methods? Wiseman (1949) concluded that the Devon method, a system of multiple marking and writing to develop reliability, is one way. Unfortunately, certain British EFL tests, amongst others, do not use multiple marking methods because of the cost involved and instead use a system of moderation to check the accuracy of examiners by sampling their scoring. It is therefore almost impossible to evaluate the reliability of these examinations (Hamp-Lyons, 2003). To ensure some reliability, each test developer should have awareness of all the variables he/she is dealing with and limit them appropriately from the start.

4. Interest

Learner interests are very frequently negated, often simply because teachers deal with large numbers of students and cannot possibly cater for all their individual needs. However, studies, such as those by Hinkel (2002) and Hamp-Lyons & Kroll (1996) have shown that learner interests may have a direct correlation to test results. Hamp-Lyons & Kroll (1996) explained that students' writing abilities are going to be reduced to a smaller score range than may possibly happen if students could discover their own level by writing on comfortable topics. The reason for this is simple: the more familiar or interesting a topic, the more cognitive schematic knowledge is available for application. Hinkel (2009) also

found that the choice of topic has a direct impact on the occurrence of modal verbs in L2 writing tasks. It may be impossible to accommodate all student interests but, as can be seen, it is important that test designers take into consideration the assumed background knowledge and experience of test-takers to ensure the fair measurement of students' skill capabilities. On the other hand, examiner interests should also be taken into consideration. Candidates should be engaged enough in the task to find something to say, but examiners should be sufficiently interested in the topic in order to read many essays on that given topic (Weigle, 2002). This dilemma may be circumvented by offering a limited range of topics so that students have a choice, though this may influence result reliability. Ultimately, the test designer needs to consider all these factors and decide on fair, valid and reliable tasks and measurements.

Conclusion

In this paper, the factors that may influence the direct testing of students' writing abilities were considered. The importance of these factors as minimum requirements during the test construction phase was especially emphasised. Some of these factors include clarity, validity, reliability and interest. Clarity was shown to be essential as a means for candidates to understand what is required of them quickly and easily. Next, validity was explained as a test measurement and it was concluded that a test is valid when it accurately measures what it set out to assess in the first place. Four types of validity were identified and discussed, namely face validity, content validity, criterion validity and construct validity. Thereafter, reliability was expounded. This refers to the predisposition of test scores to be the same from one test occasion to the next. Ways in which this may be ensured were also illustrated. Finally, learner and teacher interests and its correlation to test results was discussed, concluding that test designers need

application of scores is that if test scores are misinterpreted, and thus erroneously applied, the test is essentially invalid (Messick, 1994). Douglas (2000) provided the following useful example: imagine that we want to test the English communication abilities of a group of medical doctors, but we do not include a speaking component in the test. This will mean that the test results cannot be interpreted or used as information regarding successful or unsuccessful doctor communication with staff or patients.

Two construct validity factors that help to guide test construction and ensure test validity are construct representation and construct relevance. Messick (1996) explained that construct representation implies that tests not only measure the actual abilities of learners but that the tests themselves can be measured in terms of authenticity and directness. Construct relevance, on the other hand, implies that an exam tests only those skills which are part of the construct and nothing outside of it. Finally, another important factor that determines the ultimate validity of tests is that they should be unbiased and fair, and should never be used to encourage unjust treatment or harmful learning (Hamp-Lyons, 1989). Naturally, it is not the tests themselves that are harmful or beneficial to students, but the ways in which tests and particularly test results are used, both in interpretation and application (Messick, 1996; Wall, 1998). Since many of these tests are arranged by international agencies that are relatively influential, those who manage the agencies and are employed by them should take this authority and responsibility very seriously (Hamp-Lyons, 2000).

3. Reliability

Human beings do not behave in exactly the same way on every occasion, even when the circumstances seem identical (Hughes, 2003), though this is exactly what reliability implies:

the propensity of test scores to remain consistent from one test occasion to the next. This is not easily achieved as there are numerous factors to be taken into consideration during the test construction phase which may include issues such as test type, question possibilities, general test level and so on. These variables were overcome by the implementation of the multiple-choice test, also known as the objective test. However, a question remains: how does a test designer establish reliability in direct testing methods? In these cases, reliability corresponds with rating procedures. In other words, markers need to have consistent grading measurements to ensure reliable test results. Nowadays, scholars easily reach 75% and higher reliability by, for example, using more than one marker (usually one internal and one external) and by reaching consensus between the score assigned to a given piece of real writing by the two or more markers. This, together with a strong validity test constitution, is perfectly acceptable in education contexts, though what should be acknowledged is that the scores might differ for individuals as human emotional and physical states also have an impact on final test results. Thus, test results may differ if they were taken a day before or a day after (Hamp-Lyons, 2003).

A related aspect that needs to be given consideration is the complexity of a writing activity. There are certain problems associated with the expectation of reliability when viewed from the point of writing rather than from the point of rating. As was implied at the end of the previous paragraph, the same person will not write equally well on every occasion, nor will a person write equally well on diverse topics. Differences in ratings of work done by the same person therefore do not necessarily mean that the scores are unreliable. In fact, the scores may be reliably reflecting differences in

spell well are also competent in grammar. This is of course not necessarily the case and as such clear specifications about the nature of the relationship between spelling and grammar usage should be built into the test from the start.

2.3 Criterion validity

Criterion validity refers to the degree to which results on a specific test agree with those offered by certain independent parties, which have been found to be highly dependable and a reliable measurement of test takers' abilities. Consequently, the independent assessment may be used as a criterion measure against which a specific test is validated (Hughes, 2003).

Criterion validity is further sub-divided into two kinds of validity, namely concurrent validity and predictive validity. Concurrent validity is created when the test and the criterion are measured at about the same time, while predictive validity concentrates on the degree to which a test may predict test takers' future abilities (Fulcher & Davidson, 2007). Anastasi (1986) explained that predictive validation is really the predictive utility of a test while concurrent validation is the diagnostic utility of test. In other words, what do test results predict about future individual and group performance, and what do test scores reveal about individual and group problem areas?

This is an especially useful tool when dealing with L2 learners as their progress may be monitored against the predictions of teachers. If there is a great discrepancy, the tests may become a useful diagnostic tool and may show up the lack of expertise of the teacher in question, a challenging area of learning for the students or an oversight in the test design.

2.4 Construct validity

Though the contribution and reliability of the traditional face- content- criterion validities cannot be emphasised enough, it is important to understand that the construction of a good

test requires more than this. Construct validity aims to assess some kind of human response. A test therefore has construct validity only if and when it can appropriately reflect the reality of behavior in the area being tested (Hamp-Lyons, 2003). Moreover, the empirical data should reflect the behavioral consistencies of constructs (Anastasi, 1986). These identified constructs may be quite abstract, such as writing ability, but though these kinds of capabilities are difficult to observe, they can and must be assessed by recognising certain behaviors identified with the constructs. Thus, a test is considered to be valid and reliable when it correctly captures the behavioral information associated with a specific construct, but there is a difference between a reliable test and construct validity. For example, if you shoot three arrows at the same target and they all hit the same spot, your objective can be said to be reliable. However, if the same three arrows hit the edge of the goal, your objective can be said to be reliable but ineffectual. Now presume one arrow hits the bull's-eye and the other two shoot into different directions. This means that there is some degree of success or efficacy but no reliability. Lastly, presume that all the arrows hit the bull's-eye. This implies that there is both efficacy and reliability. Construct validity requires this kind of precision: hitting the right spot most of the time (Hamp-Lyons, 2003).

Most test developers have at least some idea about the constructs they want to test and once the test has been set and taken, the scores need to be interpreted. This is where validity generalisation may be useful. Validity generalisation is when the scores generated by a test make sense even to those who have not been given any descriptive information about the test and test results. It is thus useful in that it ensures that the general meaning of a test score will not be misinterpreted to the disadvantage of learners. The most important aspect of test validity in the interpretation and

allows them a better chance to apply their knowledge (Hamp-Lyons & Kroll, 1996).

2. Validity

Validity refers to everything that contributes to our understanding of what the test measures (Anastasi, 1986) and it is assumed that a test is valid once it measures accurately what it sets out to assess (Hughes, 1989). There is, however, also a need for empirical evidence to claim that a test has validity. In the mid 1950s a great effort was made to organise test construction and bring some measure of consistency to it. This resulted in the classification of validity which is used as an empirical basis and has been described as four types of validity by researchers, namely face validity, content validity, criterion validity and construct validity. Some researchers, like Cronbach & Meehl (1955), referred to only three types and disregard face validity and still other researchers, like Anastasi (1986), assumed that construct validity is actually a comprehensive approach that includes the other recognised validation procedures. This debate is beyond the scope of this paper, however, and here we will simply explain the four identified types of validity that should be considered during the test construction phase.

2.1 Face validity

Face validity, according to Nevo (1985), is a characteristic of tests that can be validly and reliably measured and presented. In other words, the assessment developed needs to be valid at face value. This means that the test designer needs to set out what he/she aims to measure in the specific assessment and then do so accordingly. This is a vital step in the design phase even though it is not a scientific notion and does not provide evidence for construct validity. However, if, for example, an examination does not have face validity, it might not be accepted by teachers or test takers, and even if it is used, the response from

students may not reflect their true abilities (Hughes, 2003).

2.2 Content validity

Content validity refers to the relevance of the content as well as content coverage which Anastasi (1986) described as domain specifications and domain representativeness, respectively. It is generally accepted that a test has content validity when its content comprises a representative sample of the language skills with which it is implied to be concerned. Moreover, a test is more likely to have content validity if the test designer first obtained a proper sample of the relevant structures. These relevant structures may rely on the objective of the test. Consequently, to judge whether or not a test has content validity, the structures need to be specified. Such specification ought to be made at an early stage in test construction. It may not be expected that everything in the specification will emerge in the test but it will at least create a foundation for the test designer from which to make principled choices of factors to include in the test (Hughes, 2003).

Content validity, however, has an inherent duality. On the one hand, the greater a test's content validity, the more accurate the assessment. On the other hand, content validity usually concentrates on elements that are easy to test rather than elements that are important to test. In other words, elements that do not appear in a specific assessment might point to the fact that those elements were also overlooked in teaching, and therefore learning. A test designer can avoid this pitfall by writing test specifications for the assessment and by ensuring that the test content fairly reflects the foregone learning (Hughes, 2003). For example, a test designer may want to analyse the relationship between spelling and grammar. If the emphasis is incorrectly placed on the spelling element of the test, the conclusion drawn may be that learners who

influence of background knowledge and experience of learners, especially L2 learners, on test results.

In order to comprehensively address the factors that need to be considered during task preparation for the direct testing of writing ability, this paper has been organised in the following way: at the outset, the role of clarity is investigated and described, after which some of the validities are outlined and discussed in the next section. Then, the focus is on reliability and finally, the role of interest is examined. The following section is thus a summary of the relevant research and a discussion of what the research indicates in terms of the effects that these factors may have on L2 learners. Finally, the findings are summarised and conclusions are drawn about the factors that need to be taken into consideration during the planning and development phases of writing assessments.

The influential factors during task preparation for direct tests of writing ability

As mentioned earlier, there are at least four factors that need to be taken into consideration throughout the preparation of writing tests, namely clarity, validity, reliability and interest. In the following sections these factors will be further discussed and illustrated.

1. Clarity

Clarity refers to all the issues related to ensuring that the meaning is clear and accessible (in terms of decoding) for learners. Hamp-Lyons & Kroll (1996) set out a table describing design guidelines in their paper which deals with some of the clarity issues that need to be taken into account. These include content, language, tasks and rhetorical specifications and will be briefly discussed hereafter.

1.1 Content

Content refers to the actual subject matter that the test is based on. When dealing with L2 learners, their varying backgrounds should be

taken into account so that the content is culturally accessible (Hamp-Lyons & Kroll, 1996). If students living in rural areas or villages, for example, are given a piece on city life that deals with concerns unknown to them, or assumes a degree of background knowledge and experience that these students do not have, it may affect the results negatively and prevent students from demonstrating their true writing capabilities. Other content factors which may influence the result are the use of language, the chosen tasks and rhetorical specifications. These will now be further elucidated.

1.2 Language

All language instructions should be clear, explicit and appropriate in terms of linguistic complexity. They should, furthermore, be culturally accessible and easy to interpret (Hamp-Lyons & Kroll, 1996), thus ensuring that students are given the opportunity to adequately demonstrate their writing skills.

1.3 Tasks

Hamp-Lyons & Kroll (1996) write that the most concrete aspect of the academic writing situation, that which as test developers we can most hope to identify and respond meaningfully to, is the range and nature of academic writing tasks. This includes the types, variation and frequency of tasks which should be determined in terms of appropriateness to the students in question. Failure to identify the needs of students may result in poorer students being unable to decode the tasks and therefore unable to respond correctly or at all.

1.4 Rhetorical specifications

Rhetorical specifications simply means that all instructions should be unambiguous and that sufficient prompts should be provided, thus creating a system of association between content, language and tasks. In other words, if phrases from the text, for example, are repeated in questions, it creates a link for students in their thinking and in so doing

The Influential Factors on the Direct Testing of Second Language Learners Writing Ability

Anwar Saad Aljadani

Abstract. this article provides an overview of certain influential factors on the direct writing test in second language acquisition. It presents the importance of some factors which are possible to have an influence on the direct testing of students' writing abilities. Some of these factors are clarity, validity, reliability and interest. Firstly, clarity is understood as a means for second language learners to get the requirements of them quickly and easily. Secondly, validity was known as a test measurement and it was concluded that a test is valid when it accurately measures what it set out to assess in the first place. Four types of validity were included, namely face validity, content validity, criterion validity and construct validity. Thereafter, reliability which refers to the predisposition of test scores to be the same from one test occasion to the next. Finally, learner and teacher interest and its correlation to test results was discussed, concluding that test designers need to weigh up all the variety of factors during the test construction phase and choose those that will result in a fair and valid test that is clear, reliable and interesting for all parties concerned.

Keywords: writing test, writing ability, clarity, validity, reliability and second language learners.

Introduction

Much has been written regarding language testing, particularly writing performance and the factors that may affect it. However, there is another step to be considered, namely the preparation and setting of such writing tests and which factors should be considered to optimise the writing environment for learners without compromising faculty expectations. Four of the factors that have been emphasised in research are clarity, the validity of a writing assessment, the reliability of a writing assessment and the interests of learners and teachers. To ensure a better understanding of these factors they need to be briefly defined here, though they will be elucidated in the sections that follow.

The first of the four factors, namely clarity, is related a wide range of issues such as content, language, tasks and rhetorical specifications (Hamp-Lyons & Kroll, 1996) which need to be taken into account during the design phase of writing assessments to allow both poorer and more competent students to demonstrate the range of their skills. The second factor, namely validity, is defined by Huot (1996) who wrote that in order for a writing assessment to be valid, it needs to have a strong theoretical basis, which clearly describes the constructs that the measurement should assess. The third factor, reliability, refers to decisions about scoring procedures (Hamp-Lyons & Kroll, 1996) and the fourth factor, interest, refers to both the type of topic and the potential