Apples – Journal of Applied Language Studies
Vol. 11, 4, 2017, 1–9

Discussion Note

Test usefulness of IELTS writing test tasks

Neng Priyanti, University of Pelita Harapan

This analysis was conducted to find out the extent to which IELTS writing test tasks are useful for its intended purpose. The indicators of test usefulness used in this study are those of Bachman and Palmer (1996): construct validity, reliability, authenticity, interactiveness, impact and practicality. This framework is taken as a viewpoint as it offers a comprehensive framework of test usefulness that is specific to that of language assessment. The result of the analysis shows that the IELTS writing test tasks place more concern on reliability and practicality. While not necessarily neglecting the other indicators, the nature of the test as a large-scale test reduces its overall usefulness. The predictive value remains unclear as its construct validity is still under-represented. The authenticity is problematic as the tasks correspond to only small parts of real tasks in universities in terms of their length, types of tasks, time allotment, and range of topics offered.

Keywords: authenticity, impact, interactiveness, reliability, validity, practicality, test-usefulness

1 Introduction

A language test is of high quality if it is useful to its intended purposes. These purposes are different depending on the types of test being administered, target population, testing situation, context and availability of resources. A classroom language test, for example, may serve as a means to measure how much test takers have learned throughout the course. As its name, this test has specific objectives, specific test taker population, and specific materials to be tested to students. This test also allows feedback on test results to be immediately given and remedial instruction to be decided and taken accordingly. This test, therefore, makes it possible test takers to re-take the test and hence make continuous improvement in terms of learning. Unlike the classroom language test, the purpose of a standardized language proficiency test such as IELTS is to measure one's general language ability. To be specific, it is to assess the language ability of those test takers who plan to study or work in institutions or places where English is used as a medium of instruction or communication. The IELTS offers two tests, the IELTS General Training and IELTS Academic. The former is for test takers who want to live or migrate to an English speaking

Corresponding author's email: nengpriyanti@gmail.com
ISSN: 1457-9863
Publisher: Centre for Applied Language Studies
University of Jyväskylä
© 2017: The authors
http://apples.jyu.fi
http://dx.doi.org/10.17011/apples/urn.201708073427

country. The latter is for test takers who want to study in universities or higher education where English is used as the medium of instruction. Like other standardized language proficiency tests, this test has been standardized for worldwide use, trying to accommodate the needs of many universities for some common ground of English language proficiency for university acceptance around the world. This can be seen through the growing number of universities which have made the IELTS test score as one of their entry requirements for university admission.

Even though this test has been widely used, there has not been any consensus on the extent to which this test is useful for its intended purposes. In fact, over the last three decades researches investigating the predictive validity of IELTS have found considerably inconsistent results. Some earlier studies, for example, found that there is a little correlation between test takers' score and test takers' later academic attainment (Davies, 1988; Dooey & Oliver, 2002; Kerstjens & Nery, 2000). Other findings such as that of Feast (2002) and Paul (2006), however, state that there is a significant correlation between IELTS score and students' Grade Point Average (GPA). Despite these conflicting results, Oliver, Vanderfold and Grote (2012) argue that this test, when compared to other available English language proficiency tests, is still the best predictor of test takers' later academic attainment.

In addition, most researches on the IELTS test are somewhat treated as a different entity that does not investigate its usefulness of its intended purposes. Some of these examples are; Allen (2016) who investigated the washback of IELTS test to the test takers in Japanese universities; Moore and Morton (2005) who examined the authenticity of IELTS writing tests by comparing the IELTS writing tasks to that of writing tasks in universities; Veerappan & Sulaiman (2012) who investigated the inter-rater reliability of IELTS writing tests. To date, there are no studies on the overal usefulness of IELTS test.

For that reason, this paper will analyze the overal usefulness of IELTS academic writing tests using the framework offered by Bachman and Palmer (1996). This model is taken as a viewpoint on the context of IELTs as the IELTS is a high-stake test, meaning that the score is used to make informed important decisions about test takers' future education (e.g., admission to university). As the significance of a test relies on its score interpretability (Weigle, 2002), an IELTS test therefore should be useful for its intended purpose. In other words, if a test is not useful for the intended purpose then the score interpretation or prediction will be misleading and thus harmful to test takers or other related stakeholders. Another reason why test usefulness is used is that this model considers the complexities of language assessment and covers the essential elements of it, and thus offers a comprehensive framework of language test usefulness. This paper will focus on analyzing the IELTS writing tests because these tests are easily accessible.

2 Construct validity

Construct validity refers to 'the meaningfulness and appropriateness of the interpretation that [are] made on the basis of test score' (Bachman & Palmer, 1996, p. 21). For a relevant and objective interpretation, Weigle (2002) argues that a test then should be developed with specific and detailed constructs of language ability, skills, or structures it is envisioned to measure. In the same vein, Hughes (2003) states that such specification is of critical importance in test development, as it provides a means for construct validation. Construct validity consists of two aspects – content and criterion validity (Bachman & Palmer, 1996). Content validity refers to the extent of test representativeness to its possible important content (Hughes, 2003). Two years later, Weir (2005) critically assimilated content and context validity, by which she argued that the content representativeness should be seen from the use of language in its contexts.

Despite IELTS claims on its content and context validity, IELTS writing task 1 and 2 may, to certain degree, not constitute a representative sample of required real-academic language skills or structures given the wide-range of educational backgrounds of test taker population. Although the discourse modes (descriptive and essay writing) are similar to those of university tasks; the length, topics, and time allotment are more to public discourses (Moore & Morton, 2005). In addition to that, level of linguistic demands of particular courses are also varied, law courses are much more linguistically demanding compared to that of pure Mathematics, for instance. As such, those test types are, to a certain degree, valid for particular courses but not for others. In fact, it is the nature of IELTS as a large-scale test that contributes to problematic content validation due to its difficulties in catering test-takers' various needs (Weigle, 2002). Therefore, this 'construct underrepresentation' (Messick, 1992, as cited in Weir, 2005) of areas such as discussed above indicates the lack of content validity.

The lack of content validity leads to lack of predictive validity. By predictive validity, Hughes (2003) refers to the extent of test-score predictability towards examinees' future performance. One's writing ability in IELTS writing timed-impromptu tasks 2 with irrelevant word-length and general topic to that of real academic tasks may not make a fair interpretation, and therefore may fail to predict candidates' performance in real academic setting. White (1994) supports this by arguing that there is a difference in terms of performance when examinees are given topics of interests or of discipline compared to more general ones.

3 Reliability

Reliability simply means the consistency measurement of test-scores (Bachman & Palmer, 1996). In other words, test is considered reliable if it can be trusted to generate almost similar results on various administration times or by different raters (Weigle, 2002). To put it simply, if the result of one test taken by one examinee within a day or two-gap interval is hugely different, the test cannot be reliable. Therefore, inferences from such scores cannot be trusted. This can be problematic for high-stake tests such as that IELTS because errors in inferences can cost a rejection of further study.

There are two types of reliability – test reliability and scoring reliability (Bachman & Palmer, 1996). The former refers to within-test variables, namely test instruction, discource channel (e.g. computer mediated or paper based) and test items that could threat test reliability. The later on the other hand refers to external variables such as marking process. The nature of writing tasks place a threat to test reliability due to its degree of subjective judgement. In other words, replicability of scores of writing tasks cannot be guaranteed (Hughes, 2003) particularly in the context of second/foreign language users which are highly influenced by the native language discoursal systems as well as by culture. For example, English discoursal system is considered to be a 'straight line', most Asians such as Indonesia is using 'inward-pointing spiral' (Weigle, 2002, p. 22). In the case of IELTS where candidates are expected to be studying within an English discoursal system, their future academic survival would depend on adopting such writing systems. Accordingly, expecting students to follow an English discoursal system is contemporarily believed to be valied for the test takers (Uysal, 2009).

To mitigate threats to scoring reliability, scoring rubrics or detailed descpritive scoring criteria are needed. Raters rely on this to minimize different interpretation of examinees' written products. In fact, IELTS use of analytic scoring rubric is one of its reliability assurance conducts. According to Weigle (2002), analytic scoring rubric can highly enhance test reliability. While rubrics could minimize the reliability threat, variables such as raters' educational and cultural background is of critical influence in scoring. Some studies have proven that raters' different educational background tend to use different benchmarks to score writing tests (Sweedler-Brown, 1993). In a similar vein, raters' familiarity with examinees' rhetorical patterns is evidently shown to affect level of leniance of scoring (Kobayashi & Rinnert, 1999). Even though these two factors could be minimized through continuous rater trainings as what IELTS does, this cannot be entirely obliterated (Weigle, 1998). To minimize such subjectivity that examineers possibly bring in IELTS, double marking standard is applied.

4 Authenticity

Given the main purpose of any language test is to make predictions and generalization about one's language ability in real world, test tasks should be authentic. By authentic, it means how test tasks 'correspond in demonstrable ways to language use in non-test situations' (Bachman & Palmer, 1996, p. 9). As such, when one task is not in congruence with what happens in real life, it is therefore unauthentic. However, authentic tasks may be problematic if transferred in testing situation, as it may conflict with aspect of reliability and practicality (Sanchez, 2004). For example, a graduate student might have to write a 4000-5000-word essay in two weeks. If this task is transferred in testing situation, there is no guarantee that the essay produced by test takers will be reliable, and it is definitely not practical. Acknowledging such problems in language teaching contexts, Nunan (2004) argues that test tasks should approximate to that of real-life tasks.

The nature of IELTS as a large-scale test places it in a difficult point to provide authentic test types for such a huge test taker population (Weigle, 2002). IELTS writing tasks 1 and 2 discourse modes (a descriptive writing and an essay) are similar to that of university tasks, the topics and length of words on the other hand is dissimilar to examinees' future real academic requirement as they might be required to write a discipline-related essay of more than just 250 words (Knoch & Elder, 2010). In addition, task 1 provides non-verbal stimulus, which is related to university tasks, task 2 is a timed-impromptu task. This contrasts to real academic world as academic writings are mostly scheduled in a more flexible time frame and writers are allowed or even suggested to refer to and cite other sources rather than depending on own ideas. Writers are furthermore writing about a topic within their own field of discipline and have prior knowledge, and therefore are well equipped in terms of ideas.

5 Interactivity

Closely linked to authencity, interactivity is defined as 'the extent and types of involment of test takers' individual characteristics in accomplishing a test task' (Bachman & Palmer, 1996, p. 25). In other words, test is interactive if there is an engagement of language knowledge, strategic, and metacognitive strategies which are unique to each individual. This is of importance because it provides a connection between language ability and real language use (Weigle, 2002). This view challenges the traditional or (probably) current belief that 'ELT in schools was, and still mainly is, a matter of teaching the form of English as knowledge' (Li, 1990).

IELTS test is considered to be relatively interactive by involving language knowledge and metacognitive strategies. To be specific, prior to writing, examinees are required to mentally set their writing goals, assessing different features of tasks, and thus planning on how to execute writing by selecting appropriate linguitics knowledge at their disposal. These stages are parts of metacongnitive strategies pointed out by Bachman and Palmer (1996). The topics offered in IELTS test, however, are limited. While IELTS test both task 1 and task 2 engage topical knowledge by allowing test takers to include personal experience and knowledge, such tasks do not activate affective schemata by relating the topics to personal and emotional experience or by offering a liberty to test takers to choose a topic of their interest. Activating schemata is necessary in writing tasks as it can improve test takers' performance in writing (Brown, 2007).

6 Impact

Impact is generally defined as the effects of tests on learning and teaching (Hughes, 2003). Impact is a term for macro level effect such as those on society and educational system, backwash on the other hand refers to effects of learning to individuals such as examinees and examiners/teachers (Weir, 2005). Messick (1996) simply defines washback within a validity framework as an 'instance of the consequential aspect of construct validity' (p. 243). Weigle (2002) on the other hand takes a more a holistic approach by arguing that washback is a rather complex phenomenon and influenced by other test-usefulness qualities.

While the use of direct writing tests in IELTS is argued to bring positive washback to the examinees, the issue of under-representation discussed under the heading of validity seemingly leads to adverse washback on examinees due to irrelavancy of test topics to examinees' future language use goals (Weir, 2005). Futhermore, it attracts a tendency towards teaching to the test which, as many have argued, is counterproductive (Haladyna et al., 1991, as cited in Qi, 2004; Weigle, 2002). A study of washback effect (Hayes & Read, 2004) reveals that

IELTS preparation classes teach strategies to pass the test rather than developing academic language proficiency. In a similar fashion, Indonesian students are allocated six months prior to national examination to study to the test through drilling and repeated rehearsals (personal observation). Test is thus seen as a goal rather than a measurement and evaluative tool which adversely leads to negative effects to not only examinees, both also to pedagogical instructions and educational systems at large.

A few years ealier, Hughes (1988) on this matter takes a rather positive view by arguing that teaching to the test might be productive as it aligns with proper objectives of the course. In the same vein, McEwen (1995, as cited in Cheng & Curtis, 2004, p. 3) argues that 'what is assessed becomes what is valued, which becomes what is taught'. This seems to suggest that washback is bidirectional in nature. In other words, the extent whether a test has a positive or negative backwash is determined by instruction and vice versa.

7 Practicality

Practicality refers to 'the relationship between the resources that will be required for test development and administration and the resources available for these activities' (Weigle, 2002, p. 56). Bachman and Palmer (1996) argue that practicality affects test development and administration in every stage. Towards this matter, Weir (2005) contends that practicality should never threaten test construct validity. In the case of large-scale tests such as IELTS, practicality consideration seems to conflict with other qualities. For example, for the sake of practicality and accessibility to its population, IELTS academic writing topics are limited to more general ones and numbers of tasks are limited to two in a 60 minute time frame, thus, placing validity quality at threat.

8 Conclusion

Bachman and Palmer (1996), whose framework this investigation is based on, maintain that a language test will be useful if it is developed with a specific purpose, a specific test taker population and a specific language use domain. In the case of IELTS test, the test taker population is unquestionably huge, hence general. It is, in fact, for people or professionals who want to apply for higher education, be it undergraduate programs or graduate programs, or for professional registration. These test takers may apply for different programs and different field of studies. Some might apply for an undergraduate program in Computer Science; some others might apply for a graduate program in Development studies, Applied Linguistics or Food technology.

These programs offer different levels of difficulty and sophistication to their prospective students. For example, written assignments of graduate students will unquestionably require certain level of complexity to finish. When compared to the assignments of undergraduate students, master's students' assignments will evidently be deeper in terms of analysis, longer in terms of word limit and more varied and sophisticated in terms of word choice. Other than the differences in term of degrees, these different fields of study also have different language use domains. The tasks of Applied Linguistics or Social

Science students, for example, are such as making essays of, say, 2000 to 5000 words. These students might also be writing reports, discussion papers, research proposals, bibliographies, etc. They might also be required to use different registers when writing up their assignments. Those in Medical school are undoubtedly using different registers than those in Law school, for instance. All of these suggest that some disciplines are more linguistically demanding compared to other disciplines.

In the case of IELTS writing test task types which are limited to only two tasks with a word limit less than 500 words, these tasks might therefore not have a high level of authenticity. This is due to the fact that they only correspond to some part of real university tasks. Other than that, the choice of topics are also barely corresponding the real university task as university students are more likely asked to write topics or essay in relation to their subjects or disciplines, thus using different registers. The limited coverage of tasks and topics of real university tasks affect construct, content and context validity of such a test which consequently will affect the predictive value of the test. In other words, the less authentic a test is, the less the predictive value of such a test.

However, to say that this test is not useful for its intended purpose might be an overstatement as to evaluate test usefulness of a large-scale test like IELTS requires complicated and multifaceted processes. This essay has only looked into and scrutinized small part of IELTS writing test tasks and has relied on limited data and literature for its investigation. For example, to analyze the impact of the test, the test does not use primary data such as test takers' academic writing bands and their later academic achievement. Even so, conclusions can still be gained from this brief investigation, that is, catering the needs of a huge test taker can be problematic as it can threaten other essential elements of test assessment. To put it another way, one test is not enough to measure one's language ability of those test takers who come from different educational and cultural backgrounds and who have different educational purposes.

References

Allen, D. (2016). Investigating washback to the learner from the IELTS test in the Japanese tertiary context. *Language testing in Asia, 6*(7) 1–20.

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

Brown, H. D. (2007). *Teaching by principles: an interactive approach to language pedagogy.* New York: Pearson Education.

Cheng, L., & Curtis, A. (2004). Washback or backwash: A review of the impact of testing on teaching and learning. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in language testing* (pp. 3–18). New Jersey: Lawrence Erlbaum Associates Publishers.

Davies, A. (1988). Operationalizing uncertainty in language testing: An argument in favor of content validity. *Language Testing, 5*(1), 32–48.

Dooey, P., & Oliver, R. (2002). An investigation into the predictive validity of the IELTS test. *Prospect, 17*(1), 36–54.

Feast, V. (2002). The impact of IELTS scores on performance at university. *International Education Journal, 3*(4), 70–85.

Hayes, B., & Read, J. (2004). IELTS test preparation in New Zealand: Preparing students for the IELTS academic module. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 87–112). New Jersey: Lawrence Erlbaum Associates Publishers.

Hughes, A. (2003). *Testing for language teachers* (2nd edition). Cambridge: Cambridge University Press.

Kerstjens, K., & Nery, C. (2000). Predictive validity in the IELTS test. In R. Tulloh (Ed.), *IELTS Research Reports* 3 (pp. 85–108). Canberra: IELTS Australia.

Knoch, U., & Elder, C. (2010). Validity and fairness implications of varying time conditions on a diagnostic test of academic English writing proficiency. *System, 38*, 63–74.

Kobayashi, H., & Rinnert, C. (1999). Factors affecting scomposition evaluation in an EFL context: Cultural rhetorical pattern and readers' background. *Language Learning, 46* (3), 397–437.

Li, X. (1990). How powerful can a language test be? The MET in China. *Journal of Multilingual and Multicultural development, 11*(5), 393–404.

Messick, S. (1996). *Validity and Washback in language testing*. Princeton: Educational Testing Service.

Moore, T. & Morton, J. (2005) Dimensions of difference: A comparison of university writing and IELTS writing. *Journal of English for Academic Purposes, 4*(1), 43–66.

Nunan, D. (2004). *Task based language teaching.* Cambridge: Cambridge University press.

Oliver, R., Vanderford, S., & Grote, E. (2012). Evidence of English language proficiency and academic achievement of non-English speaking background students. *Higher Educational Research and Development, 31*(4), 541–555.

Paul, A. (2006). *IELTS as a predictor of academic language performance, part 2. IELTS Research Reports, 4, 1-34.* Retrieved from https://www.ielts.org/-/media/research-reports/ielts_rr_volume07_report4.ashx

Qi, L. (2004). Has a high stakes test produced the intended changes? In L. Cheng, Y. Watanabe, & A. Curtin (Eds.), *Washback in language testing: Research contexts and methods* (pp. 171–190). New Jersey: Lawrence Erbaum Associates Publisher.

Sanchez, A. (2004). The task-based approach in language teaching. *International Journal of English studies, 4*(1), 39–71.

Sweedler-Brown, C. O. (1993). ESL essay evaluation: the influence of sentence level and rhetorical features. *Journal of Second Language Writing, 2*(1), 3–17.

Uysal, H. H. (2009). A critical review of the IELTS writing test. *ELT Journal, 64*(3), 314–320.

Veerappan, V., & Sulaiman, T. (2012). A review on IELTS writing test, its test results and inter rater reliability. *Theory and Practice in Language Studies, 2*(1), 138–143.

Weigle, S. C. (1998). Using facets to model rater training effects. *Language Testing, 15*(2), 263–287.

Weigle, S. C. (2002). *Assessing writing.* Cambridge: Cambridge University Press.

Weir, C. J. (2005). *Language testing and validation: an evidence based approach.* Hampshire: Palgrave Macmillan.

White, E. M. (1994). *Teaching and assessing writing: Recent advances in understanding, evaluating and improving student performance* (2nd edition). San Francisco: Jossey-Bass.

Accepted July 4, 2017