PIONEER
ACADEMIC PUBLISHING LIMITED

Journal of Linguistics and
Communication Studies
ISSN 2958-0412
www.pioneerpublisher.com/jlcs
Volume 3 Number 1 March 2024

A Comparison and Evaluation of the TOEIC and DET Writing Test

Yiran Hou[1]

[1] Zhejiang University of Finance & Economics Dongfang College, Hangzhou, China
Correspondence: Yiran Hou, Zhejiang University of Finance & Economics Dongfang College, Hangzhou, China.

doi:10.56397/JLCS.2024.03.04

Abstract

This essay aims to compare and evaluate the writing test of two widely used English language proficiency tests: the Test of English for International Communication (TOEIC) and the Duolingo English Test (DET). Bachman and Palmer's (2010) assessment use argument (AUA) is used as a framework in this essay to compare and evaluate these two tests from three aspects: test score and reliability, validity and score interpretations, and consequences. The results show that some features of the TOEIC and DET writing tests (e.g., the DET writing test's well-designed algorithm and lack of discourse level texts, the TOEIC writing test's all direct tasks and DIF analysis used) make the two tests have some strengths and weaknesses respectively in each aspect.

Keywords: TOEIC, DET, validity, reliability, language testing

1. Introduction

The test that measures people's ability in language use is called proficiency test. Proficiency test scores, especially for English language proficiency, are required by many colleges and universities all over the world to make an admission decision and predict students' academic success (Graham, 1987). As two widely adopted English proficiency tests, the Test of English for International Communication (TOEIC) and the Duolingo English Test (DET) are compared and evaluated in terms of their writing test in this essay. The evaluation is based on contemporary views on test validation, that is, an argument-based approach. To be specific, Bachman and Palmer's (2010) assessment use argument (AUA) is used as a framework in this essay, which links test-taker performance to test scores, scores to interpretations about test-taker ability and further focuses on the decisions and consequences (Llosa, 2008; Schmidgall, 2017). The following comparison and evaluation of the TOEIC and DET writing test will focus on three aspects: test score and reliability, validity and score interpretations and consequences.

2. Score Use and Reliability

According to Huot (1990), the score that can perfectly reflect a test taker's relevant knowledge is the true score, which can be totally consistent with the same test taker's future performance in the same test. In other words, a test that provides the true score has perfect reliability. However, the true score is only an

idealized version because error is inherent in any testing measurement and conditions during a test (Huot, 1990). Kline (2005) also supports that the raw score received by any one examinee is made up of a true component and a random error component which prevent the test from having perfect reliability. One source of error comes from the raters who may be affected by many factors, such as their quality, appreciation of the writing texts, and so on (Huot, 1990). Both the TOEIC and DET uses holistic scale in their free-response writing tasks to score test takers' performance. According to Montee and Malone (2013), all criteria are considered together while evaluating the overall performance to assign a single score in holistic scales. That means the scoring may be more easily affected by other factors and then involve more error components, especially when the performance is scored individually. From this point, the holistic scales used in the free-response writing tasks have a threat to the reliability of the two writing tests. However, in the DET extended writing tasks, the test taker's performance is scored by a separate algorithm that is trained on 3,626 writing performances scored by human raters with TESOL/applied linguistics training (LaFlair & Settles, 2019). In this sense, a test takers' performance in these tasks is scored by all these professionals. According to Huot (1990), when the agreement of the raters is sufficient on the score of one writing text, holistic scoring can be accepted for the error that comes from individual raters can be minimized. Therefore, the DET writing scoring technique is relatively reliable. In contrast, the responses to the TOEIC writing test are sent to ETS's Online Scoring Network and scored by certified ETS raters (Everson & Hines, 2010). Although the TOEIC program employs multiple ways (e.g., item level scoring, calibration, benchmark responses) to monitor rater performance in order to minimize the potential human error, no one can deny the fact that one written text in TOEIC is scored by no more than two raters (Qu & Ricker-Pedley, 2013). From this point, compared with the DET, the TOEIC writing test has less reliability. However, the potential faults that may appear in the algorithm applied by the DET need to be taken into consideration.

3. Validity and Score Interpretations

One component that concerns the defensibility and fairness of interpretations made on the test score is validity (McNamara, 2000). In Hughes' view (2003), Validity refers to whether the test measures what it is supposed to measure. In order to evaluate this notion, multiple kinds of evidence are required to justify that the test score reflects the test taker's target language abilities (Bachman & Palmer, 1996). There are many kinds of validity, such as construct validity, content validity, and criterion-related validity. However, according to Hughes (2003), construct can be seen as the general one in language testing. This section will display various forms of evidence to evaluate the TOEIC and DET writing test's construct validity and talk about the relevant face validity.

3.1 Construct Validity

Before analyzing the construct validity of the two writing tests, it is necessary to talk about the domain which is the start point of identifying the language knowledge and construct of a test (Im, Shin, & Cheng, 2019). Bachman and Palmer (1996) define the TLU domain as the "situation and context in which the test taker will be using the language outside of the test itself" (p. 18). That means the TLU domain relates to what type of writing ability should be assessed in a writing test. Therefore, it is necessary for the writing test to involve related task characteristics which are in accordance with the TLU domain (Wagner, 2013). The TLU domain of the TOEIC writing test is defined as the English used in the context of everyday and workplace environments (Powers, 2020). According to ETS (1998), the test tasks and test content in the TOEIC writing test are developed from the written sample collected from various personal, public and workplace environments to reflect realistic writing tasks in these contexts, which aligns with the TLU domain and can be seen as good evidence for its construct validity. Unlike the TOEIC test, the TLU domain of the DET writing test concerns English used in a variety of settings, including the universities (LaFlair & Settles, 2019). Although its extended writing tasks are designed to elicit the test takers' full range of written language abilities, such as describing, recounting, and making an argument, other tasks (i.e., yes/no text, c-test, dictation) that also aim to assess writing ability lack the language use at the discourse level which is basic and significant in various settings, especially in university (Cushing-Weigle, 2020; Wagner & Kunnan, 2015). Due to the lack of discourse level texts, these tasks also fail to assess test takers' pragmatic and sociolinguistic

competence, which can also be related to under-representation of the construct and cause invalidity (Purpura, 2004). Besides, compared with the TOEIC writing test, all the writing tasks in the DET lack characteristics to a specific purpose of language use, which also has a negative impact on the construct validity of the DET writing test.

Both the TOEIC and DET writing test assess the test taker's performance by criterion-referenced measures. The measures provide information about desired performance implied at each level of proficiency in a writing test (Glaser, 1963). Therefore, the score received by a test user can show the extent of this individual's relevant ability, which is independent of other test takers' performance (Glaser, 1963). According to Hines (2010), the TOEIC test lists three clear claims for each of its writing tasks in its specification, and the scale score and proficiency level show a close relationship with the three levels of claims about a test taker's ability. Its proficiency level descriptors also show what a test taker at a specific level can do and cannot do in detail, which fully reflects its criterion-referenced scoring system. The DET writing test applies a more direct way to establish its criterion-referenced assessment construct, which is calibrated to the Common European Frame of Reference (CEFR), an international standard used to describe multiple levels of language ability (Brenzel & Settles, 2017; Council of Europe, 2001). The criterion-referenced measures shown in the two writing tests can be good evidence for the construct validity.

It is also noticeable that the TOEIC and DET writing test do well in controlling the construct irrelevant variance, which can be reflected in the texts of these items that are designed clear and precise to make the instructions easy to understand by the test takers (Spaan, 2007). Besides, the TOEIC team emphasizes that all the teat takers are given enough time to show their abilities in the writing test, which precludes construct-irrelevance.

3.2 Face Validity

According to Hughes (2003), a test that looks like measuring what it purports to measure has face validity. Although face validity can not offer evidence for construct validity, it concerns whether a test can be accepted or used (Hughes, 2003). Test can be distinguished between direct test and indirect test. According to Hughes (1989), direct test requires the test taker to perform the skill the test wants to measure and indirect test focuses on the abilities underlying the target skill. Direct test is popular and easy to carry out while measuring productive skills (e.g., writing skill), which can be reflected in the all the three writing tasks in the TOEIC test (i.e., write a sentence based on a picture, respond to a written request, and write an opinion essay). Under the premise of being clear about the abilities the testing tends to assess, the three tasks are relatively straightforward to create realistic conditions that can elicit the test takers' relevant behaviour (Hughes, 1989). From this point, the TOEIC writing test has relatively high face validity. Similarly, direct testing exists in the DET writing test, which is reflected by its extended writing tasks. However, some indirect tasks (i.e. yes/no text, c-test, and dictation) also contribute to the overall DET writing score. While responding to these indirect tasks, test takers can not perform in a way that indeed reflects their writing ability, and they are even not required to write anything in Yes/No (text). In this sense, compared with the TOEIC writing test, the indirect tasks make the DET writing test has lower overall face validity. In addition, a test taker may show two or more abilities in these direct tasks, which makes both the test taker and the rater unable to accurately measure the abilities shown according to the score and further influence the interpretation of the test taker's writing ability.

4. Consequences

Bachman and Palmer (2010) emphasizes that the test scores should bring about beneficial consequences to the stakeholders. On the TOEIC research website, the test claim named positive impact is corresponding to the consequence in AUA, which focuses on offering benefits to the test users and having a positive impact on English learning and teaching across the world (Schmidgall, 2017). Both the TOEIC and DET writing test attempt to promote beneficial outcomes by using criterion-referenced measures to make the test takers understand they can get success as long as the performance meets the target criteria level, which encourages test takers to learn rather than competing with others (Thomson, 2012). Besides, the TOEIC carries out differential item functioning (DIF) analysis to examine the content of tasks and test performance that relatively exhibit bias in its writing test (Im & Cheng, 2019). In contrast,

there is no information showing the DET takes this analysis.

As mentioned above, the TOEIC writing test focuses on the written language ability in everyday and workplace environment and the domain is reflected in its task format and test content. Test takers in the TOEIC test can intuitively feel that the tasks may occur in their future working, which encourages their practice and stimulates learning. In contrast, although the domain of the DET test tends to be general, wide acceptance of the DET for university admission purpose possibly causes negative consequences for the test stakeholders. The task format and text used in the DET writing test are fundamentally different from those in a typical high-education course (Wagner & Kunnan, 2015). Although the DET writing items are proved to be able to predict a test taker's performance on these classroom tasks, almost no chance of meeting similar tasks in target settings threatens the motivation of the test takers who have university admission purpose to prepare for this test (Brenzel & Settles, 2017).

5. Conclusion

Overall, considering the scoring system, the DET writing test has higher reliability than the TOEIC test due to its well-designed algorithm. In the aspects of construct validity and face validity, the DET writing test can learn much from the TOEIC test. Taking into consideration DIF analysis and the overlap between the test tasks and the real language use in target settings, the TOEIC writing test can bring about more beneficial consequences than the DET test. As widely used proficiency tests, there is no doubts that both the TOEIC and DET writing test still has room for improvement.

References

Bachman, Lyle, F., & Palmer Adrian, S. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Brenzel, J., & Settles, B. (2017). The Duolingo English Test — Design, Validity, and Value. *DET Whitepaper (Short)*.

Council of Europe. (2001). *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge University Press.

Cushing-Weigle, S. (2002). *Assessing Writing*. Cambridge: Cambridge University Press.

Educational Testing Service. (1998) *TOEIC Technical Manuel*. Princeton NJ: Educational Testing Service.

Educational Testing Service. (2007). *Examine handbook listening and reading*. Princeton NJ: Educational Testing Service. Retrieved 20 March 2013 from http://www.ets.org/Media/Tests/TOEIC/pdf/TOEIC_LR_examinee_handbook.pdf.

Everson, P., & Hines, S. (2010). How ETS scores the TOEIC® Speaking and Writing Test responses. *The research foundation for TOEIC: A compendium of studies*, 8-1.

Graham, J. G. (1987). *E*nglish language proficiency and the prediction of academic success. *TESOL quarterly*, *21*(3), 505-521.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *The American Psychologist*, *18*(8).

Hines, S. (2010). Evidence-centered design: The TOEIC® speaking and writing tests. *TOEIC compendium*, 7-1.

Hughes, A. (1989). Kinds of tests and testing. *Testing for Language Teachers*, 9-21.

Hughes, A. (2003). *Testing for language teachers*. Ernst Klett Sprachen.

Huot, B. (1990). Reliability, validity, and holistic scoring: What we know and what we need to know. *College composition and communication*, *41*(2), 201-213.

Im, G. H., Shin, D., & Cheng, L. (2019). Critical review of validation models and practices in language testing: their limitations and future directions for validation research. *Language Testing in Asia*, *9*(1), 1-26.

Im, G. H., & Cheng, L. (2019). The test of English for international communication (TOEIC®). *Language Testing*, *36*(2), 315-324.

Jafarpur, A. (1987). The short-context technique: an alternative for testing reading comprehension. *Language Testing*, 4/2, 195-220.

Kline, T. J. (2005). Classical test theory: Assumptions, equations, limitations, and item analyses. *Psychological testing: A practical approach to design and evaluation*, 91.

LaFlair, G. T., & Settles, B. (2019). Duolingo English test: Technical manual. Retrieved April, 28, 2020.

Llosa, L. (2008). Building and supporting a validity argument for a standards-based classroom assessment of English proficiency based on teacher judgments. *Educational Measurement: Issues and Practice*, *27*(3), 32-42.

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

Montee, M., & Malone, M. E. (2013). Writing scoring criteria and score reports. *The companion to language assessment*, 2, 847-859.

Powers, D. E. (2010). Validity: What does it mean for the TOEIC tests. *ETS Research.* Rep. No. TC10-01. Princeton, NJ: ETS.

Purpura, J. (2004). *Assessing grammar*. Cambridge, UK: Cambridge University Press.

Qu, Y., & Ricker-Pedley, K. L. (2013). Monitoring individual rater performance for the TOEIC speaking and writing tests. The research foundation for the TOEIC tests: A compendium of studies, 2, 9-1.

Schmidgall, J. E. (2017). Articulating and evaluating validity arguments for the TOEIC® tests. *ETS Research Report Series*, *2017*(1), 1-9.

Spaan, M. (2007). Evolution of a test item. *Language Assessment Quarterly*, *4*(3), 279-293.

Thomson, S. (2012). The effects of TOEIC education in South Korean universities. *Unpublished doctoral dissertation.* The University of Birmingham.

Wagner, E. (2013). Assessing listening. *The companion to language assessment*, 1, 47-63.

Wagner, E., & Kunnan, A. J. (2015). The Duolingo English test. *Language Assessment Quarterly*, *12*(3), 320-331.