Scientific Research Publishing

Open Journal of Modern Linguistics, 2024, 14, 1-7
https://www.scirp.org/journal/ojml
ISSN Online: 2164-2834
ISSN Print: 2164-2818

# Assessment of Test Validity in the Context of the Duolingo English Test

**Chenxi Chen**

Languages, Cultures and Linguistics, University of Southampton, Southampton, UK
Email: gwychenchenxi@163.com

**How to cite this paper:** Chen, C. X. (2024). Assessment of Test Validity in the Context of the Duolingo English Test. *Open Journal of Modern Linguistics, 14,* 1-7.
https://doi.org/10.4236/ojml.2024.141001

**Received:** December 2, 2023
**Accepted:** January 26, 2024
**Published:** January 29, 2024



## Abstract

This study aims to explore four dimensions of test validity: face validity, construct validity, consequential validity and criterion validity in the context of the Duolingo English Test. Through descriptive research, it brings about some inspiration for test takers to consider how to choose an appropriate test.

## Keywords

Assessment, Test Validity, Duolingo English Test

## 1. Introduction

Due to the restriction of individual activities during the COVID-19 pandemic, many online tests have grown tremendously. As Idnani et al. (2021) stated that online tests play a crucial role in distance education, especially in an unforeseen state like the COVID-19 lockdown. Considering the minor differences between online tests and paper-and-pencil tests (Clark et al., 2020; Prisacari et al., 2017), assessing the validity of online testing can support the proposal of using effective alternatives in limited conditions.

Furthermore, sometimes tests like IELTS were cancelled for the time being because of unforeseen lockdowns in the city. The test-takers who have received a conditional offer to study in countries like the UK are very anxious because of the approaching deadline for turning in a language score certificate. To choose an alternative test that can be taken spontaneously becomes an urgent task. Therefore, selecting a test that is fit for purpose, that is, using validity to identify a good test has become an unavoidable consideration. Contrary to other expensive and restricted access language tests, the Duolingo English Test (called DET for short) relies on its convenient accessibility, affordable, and rapid score reporting advantages to appeal to a mass audience (Brenzel & Settles, 2017). To be

more specific, there are three main reasons for one to choose DET. Firstly it can be taken at home or office with an equipment like camera to supervise the examination of circumstances on the spot to ensure fair and open implementation of the test. Secondly, it is much cheaper to take a DET than IELTS or TOEFL. Last but not least, DET reports the score about three days after the test.

This article uses Messick's (1989, cited in Messick, 1996) definition of test validity that the validity of a test could be regarded as a comprehensive inference. In fact, it is necessary to explore four dimensions of test validity: face validity, construct validity, consequential validity and criterion validity. Therefore, this article explores the link between these sub-validities and DET through meta-analysis research, aiming to help test takers consider whether it is an appropriate test or not.

## 2. The Key Concept of the DET Validity

### 2.1. The Definition of Test Validity

With the shift in educational concepts, a wide range of students intend to study abroad and thus need to select an appropriate language test that can prove the levels of their language proficiency. There is no doubt that students tend to select a test that has powerful validity or is fit for the test purpose. Validity is an integrated evaluative measure of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores and other modes of assessment (Messick, 1989, cited in Messick, 1996). In brief, test validity refers to a comprehensive inference evaluated by appropriate and adequate test scores. Brenzel and Settles (2017) in their research state that the DET has a strong correlation with other popular language tests such as TOEFL and IELTS, which shows that the DET scores can be accepted for language learners and institutions. However, given the practical need to use validity in the DET, it is unavoidable to explore the classifications of test validity in depth.

### 2.2. The Classification of Test Validity

Previous empirical studies (e.g., Yao, 2023; Brenzel & Settles, 2017) have explored different sub-validities connected with the DET scores, but rarely synthesize assessing the effect of four sub-validities within the DET implementation. While DET validity in this study can be effectively evaluated by the following dimensions: face validity, construct validity, consequential validity and criterion validity.

Firstly, the face validity of a language test can be evaluated immediately by the test taker's performance. For example, if the teacher prepares a cloze test based on the latest text to examine whether students have reviewed vocabulary, this test can be considered to have face validity. Interestingly, Rubio (2005) argued that it is less rigorous to only make the judgement by face validity because it is a subjective measure of the test content. Hence, it is necessary to use other validity dimensions to support the performance of the test.

Secondly, construct validity aims to use the accumulation of evidence to support the test taker's measuring actions whether similar to the inferences. According to Teglasi (1998), the reason why construct validity plays a significant role in evaluating test validity is that it represents a shift from prediction to explanation. In a way, focusing on construct validity might support the researcher to assess the DET validity with evidence instead of subjective judgment.

Thirdly, consequential validity refers to "the appraisal of the potential and actual consequences of a test" (Reckase, 1998: p.15). It can present whether the test content matches what testers learned. Roediger and Karpicke (2006) support that setting tests helps students pay more attention to the recognition and memorization of the knowledge they have learned. However, if the assessment content is inadequate with the test takers' abilities, the test taker who passes the test may not have the required abilities. Accordingly, in the present study, consequential validity as an essential dimension of test validity is also valued.

Lastly, the study by Cronbach and Meehl (1955, cited by Shou et al., 2022) demonstrated that criterion validity indicates the degree of convergence between the test taker's score and the criterion score. However, interestingly, this is contrary to a study conducted by McDonald (2005) who argued that there is no real need for any theory to support criterion validity because if test scores could not be well correlated or are even negatively correlated with criterion scores, this test would not be the valid measurement that originates from the same concept. In some way, this statement confirms criterion validity could not be neglected in the test, also exploring the value of criterion validity still could not be thrown out. Similar to other approved language tests such as IELTS or TOEFL, testers in DET also can visually self-appraise their language proficiency with relevant language requirements. All in all, using test validity properly not only helps test takers accurately know about their language proficiency but also allows them to select the appropriate test according to their purposes.

## 3. The Analysis of the DET Validity

As the focus is on determining the validity of language tests empirically, this research mainly adopts the method of literature research for sub-validities of DET. Meta-analysis, meanwhile, is conducted in this paper since it is a rigorous approach that can compare, and even extract common results from a wide range of literature (Dörnyei, 2007).

To identify reliable representations of each sub-validities for this meta-analysis, the DET official handbook was searched. The author searched for corresponding aspects using each definition of sub-validities. Altogether, four manifestations of separate sub-validities were presented and analysed.

At first, from the Duolingo English Test official guide (Duolingo, Inc., 2021), face validity can be observed. Different from other language tests such as the IELTS or the TOEFL, the DET provides a total score along with four sub-scores as follows:

- Literacy: the test taker's English reading and writing abilities.
- Comprehension: the test taker's English reading and listening abilities.
- Conversation: the test taker's English listening and speaking abilities.
- Production: the test taker's English writing and speaking abilities.

These sub-scores demonstrate that the DET fits the goal of a language test: to assess the test taker's listening, speaking, reading and writing skills. In other words, these scores can help test takers intuitively evaluate their language strengths and weaknesses very well, thereby effectively improving them within a short time.

Secondly, verifying the construct validity of the DET should consider why the DET is a test that can satisfy the test taker's need. In other words, what is the aim of constructing the DET? As stated in the official research (Duolingo, Inc., 2021), "The DET is designed to measure integrated English reading, writing, listening and speaking skills in alignment with the Common European Framework of Reference (CEFR)." The CEFR is an international criterion for the evaluation of individual language proficiency, using a total score to indicate the test taker's comprehensive language proficiency (Settles et al., 2020). There are six grades from A1 for beginners to C2 for advanced learners, which helps test takers aware of the score that they need to achieve. In all, given DET can explain test takers' language proficiency reasonably, even can compared with other acceptable language tests, it might be viewed as an appropriate language test for the public using a professional assessment criterion.

Besides, another aspect of construct validity is that different items have their language skill requirements. The provided criteria are helpful for a test taker to answer questions logically and adequately. For instance, the speaking aural question is one of the items that assess the test taker's listening and speaking skills. A test taker is required to listen to a question three times and then answer it reasonably in a limited time. If the answer meets the key points in the questions, the test taker will be considered to not only have sufficient materials to talk about the question but also, especially, be able to understand the question's meaning with no barriers.

Concerning consequential validity, attention should be paid to the outcomes of the language test. As claimed in the official research (DET, 2021), the consequences of using scores can be used to estimate whether a test conforms to its design and purposes. Normally, the consequences of test scores are evaluated to gain some good effects for stakeholders. For example, whether teachers and test takers can accept this test or not will be considered. In some situations such as undergraduate and postgraduate admissions, DET scores can be used to prove the language proficiency of candidates. Nonetheless, unintended consequences equally deserve attention because they may result from misuse or misunderstanding. Taking an example from the official research (Duolingo, Inc., 2021), some test takers from non-English countries may often type some non-English characters and punctuation that the test system cannot recognize, thus being placed in a disadvantaged position. This kind of unfairness problem as a poten-

tial unintended consequence will be revealed when test takers have different typing or writing conventions. Therefore, a professional test should carefully evaluate the grading system and ensure that test takers will not be disturbed by unfair issues.

Lastly, exploring criterion validity in the DET should ensure the test taker's actions or responses are well correlated with the criterion answers. For example, in the extended speaking picture description section, a picture is provided for test takers to use one or more sentences to describe it in detail and accurately. Similarly, in the extended writing independent text section, test takers are required to write logically and fluently to talk about the topic. In these items, test takers should respond effectively given the offering information. Then the rating system and human scorers allow the grading rubric to rate responses (Duolingo, Inc., 2021), while the scores can intuitively observe the similarities between test takers' answers and standard ones.

## 4. Conclusion

To conclude, DET as a modern English proficiency assessment has been accepted by today's international institutions and learners. This article presents a small portion of ever-growing research that supports the validity of the DET, aiming to assess the use of different kinds of sub-validities: face validity, construct validity, consequential validity and criterion validity in the DET.

The research into the test validity of the DET has shown that the DET is an effective language test. The fact is that test takers can benefit from the convenient testing experience in an online mode. Its price is also affordable and therefore accepted by most test takers. Furthermore, DET is accepted by more and more western universities as a means to evaluate the applicants' language ability.

However, it is inevitable to recognize that there exist some practical challenges for a modern language test. The basic one is that compared with traditional language tests like the IELTS and the TOEFL, the social acceptability of the DET needs to be considered. Although individual language proficiency can be proved by the DET in some education institutions, some prominent universities such as the Ivy League demand specific language tests. Thus, it is necessary to check the admission requirements of universities before selecting the appropriate language test. Additionally, the role of standardized language tests in learners' future academic ability engendered controversy. Some international learners proposed that the templates and techniques used in language tests are hard to transfer to their academic writing. Considering sustainable development for individual education, it is crucial to keep learners' interest in professional reading after grasping fundamental language knowledge.

The validity of the DET is tentatively discussed in this study. Given that only using these four validity dimensions to evaluate the validity of the DET is insufficient, it is considerable to supplement other sub-validities in further studies,

aiming to ensure that test takers can select the language test fit for their purposes.

## Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

## References

Brenzel, J., & Settles, B. (2017). *The Duolingo English Test—Design, Validity, and Value* (pp. 1-3). DET Whitepaper (Short). https://s3.amazonaws.com/duolingo-papers/other/DET_ShortPaper.pdf

Clark, T., Callam, C., Paul, N., Stoltzfus, M., & Turner, D. (2020). Testing in the Time of Covid-19: A Sudden Transition to Unproctored Online Exams. *Journal of Chemical Education, 97,* 3413-3417. https://doi.org/10.1021/acs.jchemed.0c00546

Dörnyei, Z. (2007) *Research Methods in Applied Linguistics: Quantitative Qualitative and Mixed Methodologies*. Oxford University Press.

Duolingo, Inc. (2021) *Analysis of the Validity, Design and Development of the Duolingo English Test*. Duolingo, Inc. https://d23cwzsbkjbm45.cloudfront.net/media/resources/standards/validity.pdf

Idnani, D., Kubadia, A., Jain, Y., & Churi, P. (2021). Experience of Conducting Online Test during Covid-19 Lockdown: A Case Study of NMIMS University. *International Journal of Engineering Pedagogy, 11,* 49-63. https://doi.org/10.3991/ijep.v11i1.15215

McDonald, M. (2005). Validity, Data Sources. In *Encyclopaedia of Social Measurement* (pp. 939-948). Elsevier. https://doi.org/10.1016/B0-12-369398-5/00046-3

Messick, S. (1996). Validity and Washback in Language Testing. *ETS Research Report Series, 1,* i-18. https://doi.org/10.1002/j.2333-8504.1996.tb01695.x

Prisacari, A., Holme, T., & Danielson, J. (2017). Comparing Student Performance Using Computer and Paper-Based Tests: Results from Two Studies in General Chemistry. *Journal of Chemical Education, 94,* 1822-1830. https://doi.org/10.1021/acs.jchemed.7b00274

Reckase, M. (1998). Consequential Validity from the Test Developer's Perspective. *Educational Measurement: Issues and Practice, 17,* 13-16. https://doi.org/10.1111/j.1745-3992.1998.tb00827.x

Roediger, H., & Karpicke, J. (2006). Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention. *Psychological Science, 17,* 249-255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Rubio, D. (2005). Content Validity. In *Encyclopaedia of Social Measurement* (pp. 495-498). Elsevier. https://doi.org/10.1016/B0-12-369398-5/00397-2

Settles, B., LaFlair, G., & Hagiwara, M. (2020). Machine Learning-Driven Language Assessment. *Transactions of the Association for Computational Linguistics, 8,* 247-263. https://doi.org/10.1162/tacl_a_00310

Shou, Y., Sellbom, M., & Chen, H. (2022). Fundamentals of Measurement in Clinical Psychology. *Comprehensive Clinical Psychology, 4,* 13-35. https://doi.org/10.1016/B978-0-12-818697-8.00110-2

Teglasi, H. (1998). Assessment of Schema and Problem-Solving Strategies with Projective Techniques. *Comprehensive Clinical Psychology, 4,* 459-499. https://doi.org/10.1016/B0080-4270(73)00005-5

Yao, D. (2023). Examining the Subjective Fairness of At-Home and Online Tests: Taking Duolingo English Test as an Example. *PLOS ONE, 18,* e0291629. https://doi.org/10.1371/journal.pone.0291629