Optimizing Higher Education Learning Through Activities and Assessments

Yukiko Inoue-Smith
University of Guam, Guam

Troy McVey
University of Guam, Guam

A volume in the Advances in Higher Education and Professional Development (AHEPD) Book Series

IGI Global
PUBLISHER of TIMELY KNOWLEDGE

Chapter 10
The High Stakes Use of Language Proficiency Tests as Illusio and Pyramid Scheme:
An Evaluation of Their Social Aspects, Validity, and Reliability

Rifat Kamasak
https://orcid.org/0000-0001-8768-3569
Yeditepe University, Turkey

Mustafa Ozbilgin
Brunel University, UK

Ali Rıza Esmen
Independent Researcher, Turkey

ABSTRACT

There is a growing trend in using high stakes standardised test scores to evaluate individuals' academic and professional language proficiency. Although these tests determine the fates of millions of students and job seekers across the world, several aspects of these tests such as their design, ethical implementation, procedural fairness, and validity and reliability are questioned by many linguists. This chapter aims to evaluate the mostly criticised social and technical aspects of high stakes language tests from a pyramid scheme perspective. In order to achieve this aim, a number of empirical studies from the extant literature are reviewed, and some comments are provided in the conclusion section.

DOI: 10.4018/978-1-7998-4036-7.ch010

INTRODUCTION

Internationalisation of English language credentials as proxy for competency across many fields of work and professions has meant that the use of standardised language tests has become globally widespread. High-stakes standardised language tests such as International English Language Testing System (IELTS) and the Test of English as a Foreign Language (TOEFL) which are commonly used across the world for recruiting people, issuing certificates by recognised bodies, and allowing one to enter an education organisation (Cho & Bridgeman, 2012; Wilson, 1999; Zahedi & Shamsaee, 2012) influence the future academic and professional life of many people (Chalhoub-Deville & Deville, 2006; Deygers, 2017; Kane, 2013; Pearson, 2019). Hamid (2016) claims that institutions such as British Council, Cambridge Assessment English, IDP (International Development Program of Australian Universities) and ETS (Educational Testing Service) "are invested with enormous power to shape the destinies of millions of people globally" (p. 472). The context in which high-stakes language tests are accepted as proxy for a wide range of work and life related competencies can be framed with the Bourdieusian concept of illusio (Bourdieu & Wacquant, 1992). *Illusio* is the allure of a cultural, social, or economic game that people play and lose their ability to develop a healthy vision of the game by the act of playing it routinely. The authors of this chapter question here how the high-stakes language tests are so widely used above and beyond their original intended purpose that they constitute an illusio.

Chalhoub-Deville and Deville (2006) consider high-stakes language tests as "a linguistic threshold that enables [students] to approach academic work in English in a meaningful manner" (p. 520). Yet, several researchers (i.e., Deygers, 2017; McNamara & Ryan, 2011) who emphasise the important role of language tests in distributing social justice draw attention on ethical testing, procedural fairness, inequities or imbalances which may not always be present prior to the introduction of that test. Pearson (2019) criticises high-stakes language tests in seven aspects: "the Englishes of the test, idiosyncrasies specific to the writing modules, test fees, the interpretation of scores, test feedback, the management of challenges to results, and the retake policy" (p. 198). According to Thorpe et al. (2017), high-stakes tests create a multibillion-dollar testing industry which is rooted to another financially driven global higher-education industry. Like all other speculative industries, which grew with the hope of immediate individual gain, language test industry also displays the hallmarks of a pyramid scheme in terms of its toxic appeal, and its highly speculative promises of credentials and upward social and career mobility.

Pyramid scheme is a good metaphor for high-stakes language tests as a pyramid scheme is defined as a scheme which attracts a large number of people under often fraudulent, short term, or other unsustainable promise (Nat & Keep, 2002; Nolasco, Vaughn, & del Carmen, 2013). A pyramid scheme is sustained because people who join the game with the allure of the scheme fail to develop a healthy distance to the rules of the game. This is similar to the notion of illusio by Bourdieu and Wacquant (1992) who elaborate that when people join a game, the allure of the game prevent them from developing a healthy view and a critique of the game. High-stake languages tests are not simple business frauds. What makes the difference between a business fraud and a pyramid scheme is that the latter has a certain allure and promise that many of the participants believe in the return on their investment (Krige 2012), even when only the initial few participants could truly benefit, as it is the case with a high-stakes language test, that now has become too widely used to generate the promised returns for all participants as it did once.

Focusing on the speculative nature of the promise of the high-stakes language tests, many scholars have critiqued the promise and reality of these tests. First, the authors examine the fit of these tests for their stated purpose: their reliability and validity. High-stakes language tests are claimed to be reliable

representatives of English language proficiency of candidates, yet their validity and reliability are still examined by linguistic scholars (Jenkins, 2006a; Katalayi, 2018; Stoynoff, 2009; Yu & Richardson, 2015). Although some researchers (i.e., Fitzpatrick & Clenton, 2010; Shepard, 1993) criticise criterion-oriented (predictive), content and construct validity as the elements of "old trinitarian doctrine," the trinitarian approach is still widely used for assessment of test validity (Chapelle, Enright, & Jamieson, 2008; Fulcher & Davidson, 2007; Kane, 2013). Additionally, reliability which is defined as "the consistency of test scores across facets of the test" (Fulcher & Davidson, 2007, p. 15) was considered as a quality of a test for many years.

However, a common view which accepts reliability as a kind of validity evidence and uses reliability coefficients (i.e., Brennan, 2013; Cronbach, 2004; Lee, 2006) "to support test developers' claims of construct and content validity" (Stoynoff, 2009, p. 3) has emerged in the literature. In line, a literature review reported in the current chapter will treat reliability in a more validity integrated fashion.

Therefore, drawing on the extant literature, this chapter focuses on social aspects along with the validity measures which raised concern on justification of the high stakes use of language tests.

BACKGROUND

Do Language Tests Do What They Are Supposed to Do?

One of the ways to assess whether language tests predict a person's widely ranging competencies across all life domains, it would be important to understand what a language test accurately measures. Validity in testing refers to "discovering whether a test measures accurately what it is intended to measure" (Hughes, 1989, p. 22), or "uncovering the appropriateness of a given test or any of its component parts as a measure of what it is purposed to measure" (Henning, 1987, p. 170). Validity of a test should be determined by multiple kinds of evidence derived from different qualities of test usefulness (Fitzpatrick & Meara, 2004; Fulcher, 2015; Sercu, 2004). Therefore, using evidence from empirical studies which consider criterion-oriented (predictive), content and construct validity as distinct qualities of tests proposed by Cronbach and Meehl (1955), this chapter evaluates whether the high stakes use of standard language tests is justified empirically. In doing so, the authors also question to what extent the high-stakes language tests should be used beyond their original purpose, to be precise, to assess language proficiency, and serve as proxy for measuring life and work competencies.

SOCIAL ASPECTS OF HIGH STAKES LANGUGE TESTING

The second aspect of the promise of the high-stakes language tests go beyond an individual's competence of language. These test results today are used as markers of social standing of individuals (Au, 2013; Pearson, 2019). The unveiled assumption behind the high-stakes test assessment is that every test taker had the equivalent schooling or education quality even if not identical (Au, 2008; Bernstein, 1996; Tan, 2020). In line, Au (2013) suggests that "under the assumption that standardised tests provide fair and objective measurement of individuals, such testing seemingly held the promise that every test taker is offered a fair and equal shot at educational, social, and economic achievement" (p. 13).

Yet, educational standards and systems present a considerable number of inequalities in many countries. Certain categories of students (i.e., working class, low income, ethnic minority) have much worse educational opportunities than their counterparts (Au, 2008; Reyes, 2019). Yet, all test-takers are treated in the same way and disadvantaged students are expected to show a competitive performance in the high-stakes test assessment system. In fact, in some countries, language test results could indicate both objectively and subtly positions of social class and even professional competence (Ingram & Bayliss, 2007; Panfilova, Panfilov, & Merzon, 2015). However, critical perspectives (i.e., Au, 2013, 2016; Bağlama, 2019; Baker et al., 2010; Barrow & Rouse, 2006; Tan, 2020) on how high-stakes testing can lead to race and class inequality as well as promoting the ideology of meritocracy are more prevalent in sociolinguistics literature.

Empirical evidence that underpin the critical approaches to high-stakes language testing are abundant. Research highlights (Ali, Hamid, & Hardy, 2018; Allen, 2016; Au, 2016; Deygers, 2017; Kwon, Lee, & Shin, 2017; Reyes, 2019) that standard high-stakes language testing create disproportionately negative effects on particular groups (i.e., low-income and non-white test-takers) and nationalities (i.e., developing country citizens) generally. Amrein and Berliner (2002) found a significant correlation between the increased use of high-stakes tests and school dropout rates. The findings also revealed that African American and Latino students were twice as likely as white students to drop out of school. The study by Laird et al. (2006) which measured the relationship between the use of high-stakes tests and school dropout rates reported that students from low-income families were five times more likely to drop out than students from high-income families. Au (2013) claims that "problems like racism and class privilege are thus supposedly ameliorated through [high stakes] testing" (p. 13).

Apart from reproducing race-based and economic class-based inequalities in education, high-stakes language testing was also criticised in terms of its negative washback and pejorative effects (Angle, 2009; Lior, 2018). Tan (2020) mentions these effects as "the inordinate amount of time spent on teaching to the test, rote-memorisation, and seeking personal gain by obtaining high test scores" (p. 138).

In some cases, high-stakes language test results may even enhance a person's social and career chances, by signalling international and intercultural competence, intelligence, and cosmopolitanism (Moses & Nanna, 2007; Ramlackhan, 2020). Being competent in another language is often considered a significant credential that signals social, cultural, and economic status internationally. Yet, most of the high-stakes language tests' (i.e., IELTS and TOEFL) listening sections have a tendency of highlighting the linguistic norms of inner circle Englishes (Pearson, 2019; Uysal, 2009) which confer disadvantages to candidates who had and will have limited exposure to linguistic backgrounds associated with inner circle English norms.

Therefore, the growing critiques regarding the high-stakes language testing requires a thorough examination that includes theoretical, professional, ethical, and pragmatic considerations (Ali et al., 2018; Downing & Haladyna, 2006). High-stakes language tests are now used widely beyond their original purpose as outlined above for enhancing an individual's social and symbolic respectability across many contexts. The appeal of the high-stakes language tests for assessing social standing of an individual presents an illusio. Individuals who ascribe social value to language tests lose with the allure of these tests the ability to question what these tests were intending to measure. As more players worldwide take part, and as competence in language therefore becomes less of a scarce commodity, individuals will expect less return on their investments in the language tests. The pyramid scheme may therefore become untenable.

Validity as the Vital Element in Language Testing

Validity of a language test is associated with measurement of performance, fairness, and ethical testing (Kane, 2013; Weir & Shaw, 2005). The elements of validity that are criterion-oriented (predictive), content and construct validity as distinct qualities of tests proposed by Cronbach and Meehl (1955) are reviewed below.

Criterion-Oriented (Predictive) Validity

Criterion-oriented validity deals with the "relationship between a particular test and a criterion to which people wish to make predictions" (Fulcher & Davidson, 2007, p. 4). Logically, strength of the predictive relationship between individuals' performance on an English proficiency test and future academic performance of these individuals is considered as an aspect of validity (Cho & Bridgeman, 2012; Hamp-Lyons, 2000; Stoynoff, 2009). Empirical findings (i.e., Al-Musawi & Al-Ansari, 1999; Ayers & Quattlebaum, 1992; Cho & Bridgeman, 2012; Culpepper et al., 2019; Van Nelson, Nelson, & Malone, 2004) on criterion-oriented validity or predictive validity of high-stakes English proficiency tests are inconsistent and contradictory. In an early predictive validation of language test study, Graham (1987) reviewed eighteen studies, many of which focused on TOEFL and classified them based on their results. The researcher who found inconsistent results but did not completely repudiate the predictive validity of TOEFL related lack of consistent findings with the "complex nature of the relationship between language proficiency and academic success" (Cho & Bridgeman, 2012, p. 422) and lack of adequate data.

The predictive validity of TOEFL compared to that of the First Certificate in English (FCE) which measures four skills and explicit knowledge of grammar and vocabulary was investigated by Al-Musawi and Al-Ansari (1999) on a sample of 86 English major undergraduate students in Bahrain. The researchers used sub-scores on both high-stakes tests to examine if they are correlated with academic success measured by overall GPA and GPA in English scores of the participants through stepwise regression method. No contribution from FCE and TOEFL sub-scores was observed in the prediction of both average grades earned in overall and English courses even though both tests were entered into the analysis together.

Researchers suggested FCE as a better predictor of L2 English performance based on the weak to moderate correlations between performance and FCE scores, but Cho and Bridgemann (2012) claimed that these results were the product of the statistical method used in the study. Cho and Bridgemann (2012) also stated that the stepwise regression method suffered from a multicollinearity problem which emerged from redundant and highly correlated sub-scores of FCE and TOEFL and led to misleading findings. Similarly, the study by Van Nelson, Nelson, and Malone (2004) which was conducted on 866 graduate students in English medium instruction programs aimed to explore the predictive value of TOEFL on the GPA scores of participants. Yet, no relationship between the high-stakes test score and academic successes was reported, thus TOEFL's predictive validity was not justified in this study.

In a more recent comprehensive study, Cho and Bridgemann (2012) examined the predictive power of four high-stakes admissions-related tests; TOEFL, the Scholastic Aptitude Test (SAT), Graduate Record Examination (GRE), Graduate Management Admission Test (GMAT) on overall and discipline-specific GPA's. The study also looked at the correlations between all high-stakes tests' sub-groups by academic status and disciplines to explore if they can predict each other. Fairly small but negligible predictive validity correlations were observed between some similar sub-groups of high-stakes tests, such as combined SAT reading and writing scores and TOEFL's similar sub-group scores. The predictive validity

expressed in terms of correlations between test scores and overall GPA did not appear to be strong either. Only TOEFL scores of international students contributed 3% additional explanation to the variance of overall GPA, yet this amount was not regarded satisfactory. Besides, given many factors are involved in language performance (Qi, 2005; Stoynoff, 2009; Turner, 2006), this small amount of increase in GPA might not be attributed to a previous high-stakes test score.

Therefore, criterion-oriented (predictive) validity of high-stakes tests does not seem to be justified by the literature. Indeed, predictive validity can only be achieved for language classrooms where specified tasks and criteria are defined. However, variety in education systems, learning contexts, curriculums, teaching methods, and individuals' performances makes assessment of the predictive value of standard high-stakes tests more complicated, thus it may not always be possible to predict the impact of the test on language classrooms.

Thus, the criterion-oriented validity of high-stakes tests shows the relatively speculative nature of these tests, landing support to our pyramid scheme hypothesis. In order for these tests to become more valid, current illusory expectations from these tests should be incorporated in their redesign. Only through this approach the tests could become predictive of academic and other form of success. Alternatively, it would be a good idea to accept their limited predictive capacity.

Content Validity

Another way to assess whether high-stakes language tests presents a speculative and fraudulent claim as a pyramid scheme would be to assess their internal fairness as pyramid schemes do contain considerably high risk for and bias against new entrants, whose chances to winning diminish as the pyramid scheme becomes wider at the point of entry to the system, e.g., language competence becoming wide spread.

Content validity is defined "as any attempt to show that the content of the test is a representative sample from the domain that is to be tested" (Fulcher & Davidson, 2007, p. 6). Content validity of high-stakes testing is mainly related with fairness issues, thus it raised concerns for the potential bias in test content (Jenkins 2006b; Kane, 2013; Stoynoff, 2009). Suzuki and Daza (2004) highlight that "choosing test content that is consistent with the test-takers' needs" (p. 19) is crucially important for preparation of unbiased test content which determines fairness of the test. Kane (2013) states that content validity of language tests can be achieved "if the performance domain has been carefully specified, the domain has been systematically sampled, and the performances were evaluated appropriately" (p. 5).

However, content-based test validation can always be challenged since it includes subjectivity with regard to finding relevant, important, and interesting items to the test-taker and using specific items which represent measurement in the intended content area (Kane, 2013; Onwuegbuzie et al., 2007). The determination of appropriate content to be used in the test might include high subjectivity stemming from subject matter experts. According to Davies, Hamp-Lyons, and Kemp (2003), the claims of test bias with regard to major high-stakes English language tests have little empirical evidence, yet other researchers (i.e., Canagarajah, 2006; Jenkins 2006b; Seidlhofer, 2011) suggest that exams such as TOEFL and IELTS "privilege standard varieties of English and penalise examinees for using internationally-communicative forms of the language" (Jenkins, 2006b, p. 44).

Similarly, Suzuki, and Daza (2004) who reviewed the reading section of the Test of English for International Communication (TOEIC) contended that the context of the test neither considered cultural and linguistic differences in examinees nor did reflect the potential test-takers' real-world work context but corporate interests. In line with these suggestions, Galloway and Rose (2018) state that English

language "has been appropriated by its speakers in diverse ways" (p. 3), yet high-stakes language tests "assume examinees have acquired a variety of English that approximates the norms of Standard English" (Stoynoff, 2009, p. 5).

Under these conditions no one can guarantee that millions of non-native speakers who have not been exposed to certain varieties of English might not be disadvantaged by high-stakes standardised tests (Kamasak, Ozbilgin, & Atay, 2020). Moreover, the representativeness of content was often assessed by subject matter experts who "make judgments about the degree to which the test items matched the test objectives or specifications" (Brown, 1996, p. 233). This kind of treatment can lead to biased evaluations particularly in the writing and speaking sections of the tests where highly subjective assessment occurs. Therefore, against some evidence on the content validity of early tests based on "[the] review of the test content by subject matter experts" (Angoff, 1988, p. 22), content validity of high-stakes language tests remains unproven, particularly in light of Standard English and World Englishes debate.

Yet, again, this finding demonstrates that high-stakes language tests are not bias free. They are in fact predicated on some subjective assessment criteria, that highlights the appropriateness of our speculative pyramid scheme hypothesis for these tests. Pyramid scheme with its illusio and allure draws in a global participant who continues to take the tests without developing a critical stance about the ethnocentric and culturally monolithic biases of these tests.

Construct Validity

Construct validity is defined as "the experimental demonstration that a test is measuring the construct it claims to be measuring" (Brown, 2000, p. 9). The importance of construct validity in test validation is particularly mentioned (Brown, 2000; Cronbach, 1984; Zahedi & Shamsaee, 2012) because of the difficulty in defining a construct which might be associated with many abstract meanings and might be understood differently by test-takers (Fulcher, 1996; Fulcher & Davidson, 2007). Construct validity reflects "the correspondence between a construct and a measure taken as evidence of the construct" (Hamann et al., 2013, p. 68) or if the test can measure what it really intends to measure. Construct-based validation became a widely adopted approach in language testing and is accepted as the whole of validity from a scientific point of view because of its statistical roots (Anastasi, 1986; Cronbach, 2004; Kane, 2013; Karami, 2012; Messick, 1989).

For example, Phakiti (2008) investigated the validity of strategic competence construct of Bachman and Palmer's (1996) model. Phakiti examined the relationship between learners' long-term strategic knowledge (i.e., trait strategies) which is "knowledge of how [learners] generally perceive using a set of strategies" (Phakiti, 2008, p. 260) and actual strategy use (i.e., state strategies) to L2 reading test performance over time. Two constructs, meta-cognitive strategy and cognitive strategy constructs were operationalised in strategy use questionnaires (2 states, 2 traits). If the test involves the appropriate items which enable test-takers to use their meta-cognitive skills stored in long-term memory, then these skills should affect the use of actual meta-cognitive skills which in turn influences the use of cognitive skills and language test scores would be affected. The results of statistical analyses found that trait meta-cognitive strategy use correlated with state meta-cognitive strategy use in special contexts and state cognitive strategy use affected language test scores, thus construct validity was confirmed. Yet, the process used for construct validation brings serious concerns to the mind on how a questionnaire whose data are quantitatively treated can describe the arsenal of cognitive and meta-cognitive skills or learner's experiences (Grenfell & Macaro, 2007).

Although methods like think-aloud protocol or interview might have brought more details on what cognitive and metacognitive skills test-takers used, but these methods also possess weaknesses of subjectivity and bias as well (Kovacic, 2002; Li, 2004). Besides, relying on self-reported measures and difficulties in interpreting Likert-scale intervals might also raise questions on the accuracy of measurement (Dörnyei & Ryan, 2015).

Similar kinds of advanced statistical procedures such as factor-analysis, structural equation modelling was conducted to define the constructs and measure internal consistency of test items in high-stakes tests (i.e., Culpepper et al., 2019; O'Loughlin & Wigglesworth, 2007). From a statistical point of view, high-stakes tests seem to show rigorous results on construct validity, score dependability, and sources of bias (Cooze & Shaw, 2007; Zahedi & Shamsaee, 2012).

However, complex interactions among many influential variables might veil the real situations even if the most advanced and sophisticated statistics models were employed. Overall, research on construct validity of high-stakes tests yield contradictory results. While O'Sullivan (2005) found somewhat low reliability estimates, in particular, for the writing and speaking components of IELTS, Zahedi and Shamsaee (2012) suggest that "the construct validity of IELTS and TOEFL is widely and almost totally trusted" (p. 264). The comments of Stoynoff (2009) which highlight the lack of "theoretical and empirical evidence to support IELTS test score interpretations and use" (p. 21) demonstrate the unresolved scientific nature of high-stakes test validation.

What appears to be a robust system to a wide range of international actors high-stakes language tests remain widely speculative and the overall assessment of their validity lands support our hypothesis that these tests present a pyramid scheme and as such they are not likely to deliver the expected yield to newcomers in terms of the predicted outcomes.

SOLUTIONS AND RECOMMENDATIONS

High-stakes tests play a crucial role to assess language proficiency of individuals, thus their proposed score interpretations and uses need to be justified by validity testing. The authors demonstrate through the validity tests the extent to which high-stakes language tests could be construed as pyramid schemes. As per this requirement, a considerable amount of research was conducted to understand what norms to apply in international tests of English language ability, yet high-stakes test validity continues to be a thorny issue. However, developments in language testing are proceeding and new frameworks adopt multi-faceted assessments which use additional qualities (i.e., interactiveness, intercultural competence, and practicality) to conceptualise validity became available (Kane, 2002; Sercu, 2004).

Involvement of key stakeholders (such as test-takers) to the processes where improvements in validity conceptualisation are aimed is crucial to understand examinee needs and make sure examinees believe that the test they are taking provides them with accurate and useful information (Weir, 2005). This is also very important for ethical testing concerns and justice perception of test-takers (Deygers, 2017; Kane, 2013). The role of sophisticated statistical methods to facilitate investigations of construct validity is undeniable but it should be noted that language performance of learners is affected by the interactive relationships of many variables that are still not understood by researchers.

Given the complex nature of language proficiency assessment, it should be emphasized that various dimensions (i.e., authenticity, task difficulty, and generalisability) other than performance-based dimensions of high-skates tests should also be concerned.

FUTURE RESEARCH DIRECTIONS

In every procedure, from obtaining information to design the proposed tasks to conceptualising the validation strategy, qualitative methods—such as think-aloud protocols, observations, interviews, and focus group discussions—should be integrated with quantitative methods to ascertain the accuracy of the test. Some studies (i.e., Lee & Greene, 2007; Tsushima, 2015) which employ both quantitative and qualitative methods "in which test-takers report how they are addressing test tasks can provide more detailed analyses of the processes being applied to test tasks" (Kane, 2013, p. 28) suggest that mixed-methods can provide more accurate results than the single testing methods. Thus, as a future research direction, language proficiency studies should employ more adequate research designs which include multiple sophisticated methods.

CONCLUSION

The authors' analyses indicate that high-stakes language tests are similar to pyramid schemes. They are predicated on assumptions that (as the authors demonstrate) lack scientific support. The reasons why the international community continues to use these tests for assessment in domains of life and work, include the many interests vested in the current system, and the absence of viable alternatives. The illusio (Bourdieu & Wacquant, 1992) of the supposed yield of high-stakes language tests is sustained through the sheer size of the population that is involved in participating in the game.

The current system presents an illusio and should be amended to avoid its collapse as a pyramid scheme. Finally, theoretically informed measures of construct validity provide the most essential element in validating a test. Validation of a test cannot completely rely on content validity, if "content validity evidence is not connected to theory-based evidence" (Stoynoff, 2009, p. 35). This is because problems regarding unrepresentative test constructs and samples of the content domain can emerge (Embretson, 2007). Although the chapter explained several different types of validity, these types are all interconnected. Given the critical importance of these language tests in individuals' lives, researchers must address the lack of concrete empirical evidence of their validity. Only then will those tested, and those who test them, know whether these high-stakes tests amount to more than a speculative pyramid scheme; and if not, how to improve them.

REFERENCES

Al-Musawi, N. M., & Al-Ansari, S. H. (1999). Test of English as a foreign language and first certificate of English tests as predictors of academic success for undergraduate students at the University of Bahrain. *System*, *27*(2), 389–399. doi:10.1016/S0346-251X(99)00033-0

Ali, Md. M., Hamid, M. O., & Hardy, I. (2018). Ritualization of testing: Problematising high-stakes English-language testing in Bangladesh. *Compare: A Journal of Comparative Education*. Advance online publication. doi:10.1080/03057925.2018.1535890

Allen, D. (2016). Japanese Cram Schools and entrance exam washback. *The Asian Journal of Applied Linguistics*, *3*(1), 54–67.

Amrein, A. L., & Berliner, D. C. (2002). *An analysis of some unintended and negative consequences of high-stakes testing*. Arizona State University, Educational Policy Studies Laboratory.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology*, *37*(1), 1–15. doi:10.1146/annurev.ps.37.020186.000245

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 9–13). Lawrence Erlbaum.

Au, W. (2008). Devising inequality: A Bernsteinian analysis of high stakes testing and social reproduction in education. *British Journal of Sociology of Education*, *29*(6), 639–665. doi:10.1080/01425690802423312

Au, W. (2013). Hiding behind high-stakes testing: Meritocracy, objectivity, and inequality in U.S. education. *The International Education Journal: Comparative Perspectives*, *12*(2), 7–19.

Au, W. (2016). Meritocracy 2.0: High-stakes, standardised testing as a racial project of neoliberal multiculturalism. *Educational Policy*, *30*(1), 39–62. doi:10.1177/0895904815614916

Ayers, J. B., & Quattlebaum, R. F. (1992). TOEFL performance and success in a master's program in engineering. *Educational and Psychological Measurement*, *52*(4), 973–975. doi:10.1177/0013164492052004021

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

Bağlama, S. H. (2019). Zadie Smith's white teeth: The interpellation of the colonial subject in multicultural Britain. *Journal of Language. Literature and Culture*, *66*(2), 77–90. doi:10.1080/20512856.2019.1638007

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., & Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers*. Economic Policy Institute, Briefing Paper #278. Retrieved from https://www.epi.org/files/page/-/pdf/bp278.pdf/

Barrow, L., & Rouse, C. E. (2006). The economic value of education by race and ethnicity. *Economic Perspectives*, *30*(2), 14–27.

Bernstein, B. B. (1996). *Pedagogy, symbolic control, and identity: Theory, research, critique*. Taylor & Francis.

Bourdieu, P., & Wacquant, L. J. D. (1992). *An invitation to reflexive sociology*. University of Chicago Press.

Brennan, R. L. (2013). Commentary on validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 74–83. doi:10.1111/jedm.12001

Brown, J. D. (1996). *Testing in language programs*. Prentice Hall Regents.

Brown, J. D. (2000). Questions and answers about language testing statistics: What is construct validity? *JALT Testing & Evaluation SIG Newsletter*, *4*(2), 7–10.

Canagarajah, S. A. (2006). Changing communicative needs, revised assessment objectives: Testing English as an international language. *Language Assessment Quarterly: An International Journal*, *3*(3), 229–242. doi:10.120715434311laq0303_1

Chalhoub-Deville, M., & Deville, C. (2006). Old, borrowed, and new thoughts in second language testing. In R. L. Brennan (Ed.), *Educational measurement*. Praeger.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. Routledge.

Cho, Y., & Bridgeman, B. (2012). Relationship of TOEFL iBT® scores to academic performance: Some evidence from American universities. *Language Testing*, *29*(3), 421–442. doi:10.1177/0265532211430368

Cooze, M., & Shaw, S. (2007). Establishing the impact of reduced input and output length in FCE and CAE writing. *Research Notes*, *30*, 15–19.

Cronbach, L. J. (1984). *Essentials of psychological testing*. Harper Row.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. doi:10.1037/h0040957 PMID:13245896

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, *64*(3), 391–418. doi:10.1177/0013164404266386

Culpepper, S. A., Aguinis, H., Kern, J. L., & Millsap, R. (2019). High-stakes testing case study: A latent variable approach for assessing measurement and prediction invariance. *Psychometrika*, *84*(1), 285–309. doi:10.100711336-018-9649-2 PMID:30671788

Davies, A., Hamp-Lyons, L., & Kemp, C. (2003). Whose norms? International proficiency tests in English. *World Englishes*, *22*(4), 571–584. doi:10.1111/j.1467-971X.2003.00324.x

Deygers, B. (2017). Just testing: Applying theories of justice to high-stakes language tests. *International Journal of Applied Linguistics*, *168*(2), 143–163. doi:10.1075/itl.00001.dey

Dörnyei, Z., & Ryan, S. (2015). *The psychology of the language learner revisited*. Routledge. doi:10.4324/9781315779553

Downing, S. M., & Haladyna, T. M. (2006). *Handbook of test development*. Erlbaum.

Embretson, S. E. (2007). Construct validity: A universal validity system or just another test evaluation procedure? *Educational Researcher*, *36*(8), 449–455. doi:10.3102/0013189X07311600

Fitzpatrick, T., & Clenton, J. (2010). The challenge of validation: Assessing the performance of a test of productive vocabulary. *Language Testing*, *27*(4), 537–554. doi:10.1177/0265532209354771

Fitzpatrick, T., & Meara, P. M. (2004). Exploring the validity of a test of productive vocabulary. *Vigo International Journal of Applied Linguistics*, *1*(1), 55–74.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing*, *13*(2), 208–238. doi:10.1177/026553229601300205

Fulcher, G. (2015). *Re-examining language testing: A philosophical and social inquiry*. Routledge. doi:10.4324/9781315695518

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*. Routledge.

Galloway, N., & Rose, H. (2018). Incorporating Global Englishes into the ELT classroom. *ELT Journal*, *72*(1), 3–13. doi:10.1093/elt/ccx010

Graham, J. G. (1987). English language proficiency and the prediction of academic success. *TESOL Quarterly*, *21*(2), 505–521. doi:10.2307/3586500

Grenfell, M., & Macaro, E. (2007). Claims and critiques. In A. D. Cohen & E. Macaro (Eds.), *Language learner strategies: Thirty years of research and practice* (pp. 9–28). Oxford University Press.

Hamann, P. M., Schiemann, F., Bellora, L., & Guenther, T. W. (2013). Exploring the dimensions of organizational performance. *Organizational Research Methods*, *16*(1), 67–87. doi:10.1177/1094428112470007

Hamid, M. O. (2016). Policies of global English tests: Test-takers' perspectives on the IELTS retake policy. *Discourse (Abingdon)*, *37*(3), 472–487. doi:10.1080/01596306.2015.1061978

Hamp-Lyons, L. (2000). Social, professional, and individual responsibility in language testing. *System*, *28*(4), 579–591. doi:10.1016/S0346-251X(00)00039-7

Henning, G. (1987). *A guide to language testing: Development, evaluation, research*. Newbury House.

Hughes, A. (1989). *Testing for language teachers*. Cambridge University Press.

Ingram, D., & Bayliss, A. (2007). *IELTS as a predictor of academic language performance, Part 1*. IELTS Research Reports Volume 7. Canberra: IELTS Australia Pty Limited. Retrieved from https://www.ielts. org/teaching-and-research/research-reports/volume-07-report-3/

Jenkins, J. (2006a). The spread of EIL: A testing time for testers. *ELT Journal*, *60*(1), 42–50. doi:10.1093/elt/cci080

Jenkins, J. (2006b). Current perspectives on teaching world Englishes and English as a lingua franca. *TESOL Quarterly*, *40*(1), 157–181. doi:10.2307/40264515

Kamasak, R., Ozbilgin, M. F., & Atay, D. (2020). The cultural impact of hidden curriculum on language learners: A review and some implications for curriculum design. In A. Slapac & S. A. Coppersmith (Eds.), *Beyond language learning instruction: Transformative supports for emergent bilinguals and educators* (pp. 104–125). IGI Global Publications. doi:10.4018/978-1-7998-1962-2.ch005

Kane, M. T. (2002). Validating high stakes testing programs. *Educational Measurement: Issues and Practice*, *21*(2), 31–41.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. doi:10.1111/jedm.12000

Karami, H. (2012). The development and validation of a bilingual version of the vocabulary size test. *RELC Journal*, *43*(1), 53–67. doi:10.1177/0033688212439359

Katalayi, G. B. (2018). Can the reading construct be examined outside the reading context? An investigation of the construct validity of an English as a Foreign Language reading test. *Journal of Language Teaching and Research*, *9*(4), 685–694. doi:10.17507/jltr.0904.03

Kovacic, I. (2002). Thinking-aloud protocol—interview—text analysis. In S. Tirkkonen-Condit & R. Jääskeläinen (Eds.), *Tapping and mapping the process of translation and interpreting: Outlooks on empirical research* (pp. 97–109). John Benjamins.

Krige, D. (2012). Fields of dreams, fields of schemes: Ponzi finance and multi-level marketing in South Africa. *Africa*, *82*(1), 69–92. doi:10.1017/S0001972011000738

Kwon, S. K., Lee, M., & Shin, D. (2017). Educational assessment in the Republic of Korea: Lights and shadows of high-stakes exam-based education system. *Assessment in Education: Principles, Policy & Practice*, *24*(1), 60–77. doi:10.1080/0969594X.2015.1074540

Laird, J., Lew, S., DeBell, M., & Chapman, C. (2006). *Dropout rates in the United States: 2002 and 2003*. US Department of Education: National Center for Education Statistics.

Lee, Y. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing*, *23*(2), 131–166. doi:10.1191/0265532206lt325oa

Lee, Y., & Greene, J. (2007). The predictive validity of an ESL placement test: A mixed methods approach. *Journal of Mixed Methods Research*, *1*(4), 366–389. doi:10.1177/1558689807306148

Li, D. (2004). Trustworthiness of think-aloud protocols in the study of translation processes. *International Journal of Applied Linguistics*, *14*(3), 301–313. doi:10.1111/j.1473-4192.2004.00067.x

McNamara, T., & Ryan, K. (2011). Fairness versus justice in language testing: The place of English literacy in the Australian citizenship test. *Language Assessment Quarterly*, *8*(2), 161–178. doi:10.1080/15434303.2011.565438

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). American Council on Education and Macmillan.

Moses, M., & Nanna, M. (2007). The testing culture and the persistence of high stakes testing reforms. *Education and Culture*, *23*(1), 55–72. doi:10.1353/eac.2007.0010

Nat, P. J. V., & Keep, W. W. (2002). Marketing fraud: An approach for differentiating multilevel marketing from pyramid schemes. *Journal of Public Policy & Marketing*, *21*(1), 139–151. doi:10.1509/jppm.21.1.139.17603

Nolasco, C. A. R. I., Vaughn, M. S., & del Carmen, R. V. (2013). Revisiting the choice model of Ponzi and Pyramid schemes: Analysis of case law. *Crime, Law, and Social Change*, *60*(4), 375–400. doi:10.100710611-013-9456-8

O'Loughlin, K., & Wigglesworth, G. (2007). Investigating task design in academic writing prompts. In L. Taylor & P. Falvey (Eds.), *IELTS collected papers: Research in speaking and writing assessment* (pp. 379–419). UCLES/Cambridge University Press.

Onwuegbuzie, A. J., Witcher, A. E., Collins, K. M. T., Filer, J. D., Wiedmaier, C. D., & Moore, C. W. (2007). Students' perceptions of characteristics of effective college teachers: A validity study of a teaching evaluation form using a mixed-methods analysis. *American Educational Research Journal*, *44*(1), 113–160. doi:10.3102/0002831206298169

Panfilova, V. M., Panfilov, A. N., & Merzon, E. E. (2015). Organizational-pedagogical conditions to form the foreign competence in students with the features of linguistic giftedness. *International Education Studies*, *8*(2), 176–185. doi:10.5539/ies.v8n2p176

Pearson, S. W. (2019). Critical perspectives on the IELTS test. *ELT Journal*, *73*(2), 197–206. doi:10.1093/elt/ccz006

Phakiti, A. (2008). Construct validation of Bachman and Palmer's (1996) strategic competence model over time in EFL reading tests. *Language Testing*, *25*(2), 237–272. doi:10.1177/0265532207086783

Qi, L. (2005). Stakeholders' conflicting aims undermine the washback function of a high-stakes test. *Language Testing*, *22*(2), 142–173. doi:10.1191/0265532205lt300oa

Ramlackhan, K. (2020). Restricting social justice practises in public education: The neoliberal stronghold. In R. Papa (Ed.), *Handbook on promoting social justice in education* (pp. 193–212). Springer. doi:10.1007/978-3-030-14625-2_117

Reyes, R. (2019). To want the unwanted: Latinx English language learners on the border. In D. DeMatthews & E. Izquierdo (Eds.), *Dual language education: Teaching and leading in two languages. Language Policy,18* (pp. 77–88). Springer. doi:10.1007/978-3-030-10831-1_5

Seidlhofer, B. (2011). *Understanding English as a lingua franca*. Oxford University Press.

Sercu, L. (2004). Assessing intercultural competence: A framework for systematic test development in foreign language education and beyond. *Intercultural Education*, *15*(1), 73–89. doi:10.1080/1467598042000190004

Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammon (Ed.), *Review of Research in Education, 19* (pp. 405–450). American Educational Research Association.

Stoynoff, S. (2009). Recent developments in language assessment and the case of four large-scale tests of ESOL ability. *Language Teaching*, *42*(1), 1–40. doi:10.1017/S0261444808005399

Suzuki, M., & Daza, C. (2004). A review of the reading section of the TOEIC. *TESL Canada Journal*, *22*(1), 16–24. doi:10.18806/tesl.v22i1.163

Tan, C. (2020). Beyond high-stakes exam: A neo-Confucian educational programme and its contemporary implications. *Educational Philosophy and Theory*, *52*(2), 137–148. doi:10.1080/00131857.2019.1605901

Thorpe, A., Snell, M., Davey-Evans, S., & Talman, R. (2017). Improving the academic performance of non-native English-speaking students: The contribution of pre-sessional English language programmes. *Higher Education Quarterly*, *71*(1), 5–32. doi:10.1111/hequ.12109

Tsushima, R. (2015). Methodological diversity in language assessment research: The role of mixed methods in classroom-based language assessment studies. *International Journal of Qualitative Methods*, *10*(2), 104–121. doi:10.1177/160940691501400202

Turner, C. E. (2006). Professionalism and high-stakes tests: Teachers' perspectives when dealing with educational change introduced through provincial exams. *TESL Canada Journal*, *23*(2), 54–76. doi:10.18806/tesl.v23i2.55

Uysal, H. H. (2009). A critical review of the IELTS writing test. *ELT Journal*, *64*(3), 314–320. doi:10.1093/elt/ccp026

Van Nelson, C., Nelson, J. S., & Malone, B. G. (2004). Predicting success of international graduate students in an academic university. *College and University Journal*, *80*(1), 19–27.

Weir, C. J. (2005). *Language testing and validation*. Palgrave Macmillan. doi:10.1057/9780230514577

Weir, C. J., & Shaw, S. D. (2005). Establishing the validity of Cambridge ESOL writing tests: Towards the implementation of a socio-cognitive model for test validation. *Research Notes*, *21*, 10–14.

Wilson, K. (1999). *Validating a test designed to assess ESL proficiency at lower developmental levels* (ETS Research Report). Retrieved from https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2333-8504.1999.tb01821.x

Yu, T., & Jennifer, R. (2015). Examining reliability and validity of a Korean version of the community of inquiry instrument using exploratory and confirmatory factor analysis. *The Internet and Higher Education*, *25*(1), 45–52. doi:10.1016/j.iheduc.2014.12.004

Zahedi, K., & Shamsaee, S. (2012). Viability of construct validity of the speaking modules of international language examinations (IELTS vs. TOEFL iBT): Evidence from Iranian test-takers. *Educational Assessment, Evaluation and Accountability*, *24*(3), 263–277. doi:10.100711092-011-9137-z

ADDITIONAL READING

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, *2*(1), 1–34. doi:10.1207154343111aq0201_1

Khan, K., & McNamara, T. (2017). Citizenship, immigration laws, and language. In S. Canagarajah (Ed.), *The Routledge handbook of migration and language* (pp. 451–467). Routledge. doi:10.4324/9781315754512-26

Norris, J. (2008). *Validity evaluation in language assessment*. Peter Lang. doi:10.3726/978-3-653-01171-5

Nussbaum, M. (2002). Capabilities and social justice. *International Studies Review*, *4*(2), 123–135. doi:10.1111/1521-9488.00258

Shohamy, E. (1997). Testing methods, testing consequences: Are they ethical? Are they fair? *Language Testing*, *14*(3), 340–349. doi:10.1177/026553229701400310

Van Avermaet, P. (2009). Fortress Europe? Language policy regimes for immigration and citizenship. In G. Hogan-Brun, C. Mar-Molinero, & P. Stevenson (Eds.), *Discourses on language and integration* (pp. 15–44). John Benjamins. doi:10.1075/dapsac.33.06ave

Walters, S. F. (2012). Fairness. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 469–479). Routledge.

Zenisky, A., & Sireci, S. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education*, *15*(4), 337–362. doi:10.1207/S15324818AME1504_02

Zhang, S. (2006). Investigating the relative effects of persons, items, sections, and languages on TOEIC score dependability. *Language Testing*, *23*(3), 351–369. doi:10.1191/0265532206lt332oa

KEY TERMS AND DEFINITIONS

High Stakes Tests: A linguistic threshold that enables people to measure their English language proficiency in a meaningful manner.

Illusio: The allure of a cultural, social, or economic game that people play and lose their ability to develop a healthy vision of the game by the act of playing it routinely.

Language Proficiency Assessment: Measuring language skills of a person through testing methods.

Pyramid Scheme: A scheme which attracts many people under often fraudulent, short term, or other unsustainable promise.

Reliability: The consistency of test scores across facets of the test.

TOEFL: A high stakes test used to assess people's English language proficiency.

Trinitarian Approach: An approach that considers criterion-oriented (predictive), content, and construct validity for the assessment of test validity.

Validity: Discovering whether a test measures accurately what it is intended to measure or not.