High-Stakes Language Testing

ANTONY JOHN KUNNAN

Introduction

Definition

Whether a test is considered high stakes or not depends not on the testing instrument per se but whether the test is used as the main basis for decision making for a major career path or life changing event. In other words, when decisions based on the interpretation of test scores or evaluations (of a single test) can lead to consequences for test takers, and when these consequences radically alter test takers' major career path or life event, such tests are termed high-stakes tests. For example, in all modern societies, there are high-stakes tests, such as high school graduation tests, college undergraduate entrance tests, university graduate or professional degree entrance tests, professional licensing tests (for medical doctors, nurses, lawyers, pilots, architects, musicians, etc.), employment tests, and immigration and citizenship tests. In these tests, score interpretations provide distinctions such as pass or fail, high or low grades, competent or not competent, and so on. Such categorizations can lead to serious consequences for test takers and therefore these tests are often called high-stakes tests.

High Stakes for Whom?

It is obvious that test takers face pressure in high-stakes tests to perform well. This means they have to know the subject matter well and conditions of the test (particularly in terms of the test tasks, sections, speededness, etc.), use appropriate test-taking strategies, and respond to test tasks efficiently in the time allotted. All this is extremely important as life and career-changing decisions may be made based on a single high-stakes test score. But not all high-stakes tests turn out to be high stakes for test takers. For example, in the United States under the concept of school accountability in the No Child Left Behind Act, there is no negative impact of poor performance on the students themselves. However, school principals and teachers are under pressure to improve their students' scores annually, because if their students do not improve, school managements could be changed, teachers may lose their jobs, and accreditations may be revoked.

In other cases, not all test takers in high-stakes tests put pressure on themselves to perform well and to get the desired benefit. For example, if a test taker who takes a high school graduation test (typically considered a high-stakes test) does not care about the outcome (and therefore does not prepare or take the test seriously), then the test may not work as a high-stakes test for him or her because of the low investment in it. At the other end of the stakes spectrum is a low-stakes test such as a class quiz that is not going to have any impact on a test takers' life or career but, if a highly motivated student in school or college, or an adult in a hobby course takes such a test very seriously he or she might put himself or herself through the same pressure situation as in a high-stakes test.

The Encyclopedia of Applied Linguistics, Edited by Carol A. Chapelle.
 Published 2013 by Blackwell Publishing Ltd.
DOI: 10.1002/9781405198431.wbeal0504

High-Stakes and Standardized Tests

Many high-stakes tests are also large-scale standardized tests which are administered to large groups of test takers in centralized locations (of the paper and pencil variety) and in standardized conditions (examples, parallel forms of tests, uniform administrative measures). This is typical of high school graduation tests or college or university entrance tests, professional licensing tests, or employment tests. Such standardized tests, typically based on content and performance standards, are widely used mainly because of the convenience associated with administration, scoring, and reporting. But in some contexts, high-stakes tests are not administered on a large scale or are not standardized; examples include high-stakes tests for immigration or citizenship.

Examples of High-Stakes Tests in Schools and Universities

High-stakes tests are most common in schools and universities around the world either as school-level graduation tests (or school-leaving or exit examinations) or as university entrance tests. Three such tests are described below.

The French *Baccalauréat*

The French *baccalauréat*, often known as *le bac*, has more than a 200-year history. Established by Napoleon Bonaparte, the *baccalauréat* is a an examination taken by students in their final year of high school (known as the *lycée*) to demonstrate their achievement in school subjects so that based on the examination results students can be placed into university and the *grandes écoles*. When the *baccalauréat* was first designed, it was an oral examination of letters and humanities in Latin and Greek; today, it has evolved into a complex set of examinations of three types: the first type of exam is the *general* exam which has three series: sciences, economics and social sciences, and literature. Then there is the *baccalauréat technologique* and *baccalauréat professionnel*. The exam series is large scale and administered on the same time and day to all test takers in the country. The questions require essay-type responses and could take about 20 hours of testing time depending on the number of classes taken by the students. Successful passage of the exam series is considered completion of the French secondary education.

The CAHSEE

The California High School Exit Examination was recently introduced largely due to the federal mandate under the No Child Left Behind legislation. All high school students in California need to pass this exam in mathematics and English in order to receive a school-leaving diploma. According to the California Department of Education, the purpose and content of the CAHSEE is to improve achievement in public high schools and to ensure that students who graduate from public high schools can demonstrate grade level competency in reading, writing, and mathematics. Similar school-leaving diploma examinations are being introduced in many states in the United States.

The TOEFL

The Test of English as a Foreign Language, developed and administered by the Educational Testing Service, Princeton, is arguably the most well-known large-scale language assessment in the world. The TOEFL has become mandatory for non-American and non-Canadian non-native speakers of English applicants to undergraduate and graduate programs in United States and Canadian English-medium universities. Although the test has become

popular worldwide, the content and structure remained more or less the same over three decades. In the 1990s, a revision project began with TOEFL staff and the Committees of Examiners and Researchers with the intention of revising the TOEFL to include more communicative constructed-response tasks, direct assessments of speaking and writing, integration of skills and modes, more diagnostic information to score users, and the possibility of computer delivery. Given the high-stakes nature of the TOEFL, the revision process was deliberate and long. The Internet-based TOEFL that is administered worldwide today is the result of these revision efforts.

Examples of High-Stakes Tests in the Workplace

High-stakes tests are becoming increasing popular in the workplace. These assessments are used for screening of applicants for employment and promotion to higher levels in professions that require specific abilities. Examples of this type of assessment include assessment for civil service, health and business professionals, language teachers, court interpreters and translators, tour guides, and air traffic controllers. Here is a description of a few such tests.

Chinese Imperial Civil Service Examination

It can be argued that the earliest known high-stakes tests were the Chinese Imperial Civil Service examination system instituted in China as early as the Han Dynasty (202 BC), continued during the Ming and ending with the Qing emperors in 1905—the longest, almost continuous, examination system, at 1299 years. These examinations were supposedly created to reduce the power of the aristocracy and create a bureaucratic class that was obedient to the emperor. They were promoted and interpreted as a way of identifying merit and capability, and offering equal opportunity thus making social mobility possible. These examinations, which were held at the district, prefectural, provincial, metropolitan, and the palace levels, were in spoken Mandarin (the official spoken dialect) and written classical Chinese. The examinations tested knowledge of the classics from the *Five Classics* and the *Four Books*. Thus, classical erudition, historical knowledge, literary style, poetry, the esoteric art of calligraphy, and the infamous rigid, formulaic parallel-prose style of the eight-legged essays were considered the symbols of the Confucianized gentry-officials. However, left out from the examination pool were large classes of individuals whose parents and families did not have the means of allowing their sons to prepare for the examinations instead of working in the fields. Additionally, the use of Mandarin as the spoken language of the examinations and the use of archaic written classical Chinese privileged the Mandarin-speaking northeast but excluded other geographical regions of China.

The TOEIC

The *Test of English for International Communication*, developed and administered by Education Testing Services, Princeton, assesses everyday English skills of people working in an international environment. It traditionally consisted of listening and reading comprehension sections in a two-hour 200-item multiple-choice test. More recently, it introduced speaking and writing sections which are optional. The test is designed to reflect actual English language use in the workplace although no specialized workplace language is required to be successful in the test. Many international employers and agencies use the test for employment or promotional purposes.

Other Tests

Many current workplace related tests for civil service, health professionals, business, and tourism around the world can be considered high-stakes tests as they are sometimes the only measure of language proficiency. A few examples include the former *Occupational English Test* used for immigrant health professionals in Australia, the *Canadian English Language Benchmark Assessments for Nurses* designed to assess the communication skills of internationally educated nurses whose first language is not English, Cambridge ESOL's *Business Language Testing Service* designed to assess business communication in English, French, German, and Spanish and the *Business English Certificates* at three levels (preliminary, vantage, and higher) which are linked to the Common European Framework of Reference for Languages.

Examples of High-Stakes Tests for Immigration and Citizenship

In more recent times, high-stakes tests have begun to enter the context of immigration and citizenship. Many countries require applicants to take a standardized language assessment in the dominant language of the country they wish to immigrate to or of which they want to become a citizen. Some of the countries that have citizenship (or naturalization) tests include the Australia, Canada, Estonia, Germany, Korea, the Netherlands, United Kingdom, and the United States. A few of these tests are described here.

The US Naturalization Test

Although Naturalization Acts from 1790 onwards have controlled and restricted immigration and citizenship, it was the Immigration and Nationality Act of 1952 that enshrined both the English language and the history and government requirement. This Act required applicants for citizenship to demonstrate their ability to *speak, write, and read English and to demonstrate their knowledge of US history, principles, and form of government*.

Until the late 1980s, this requirement was managed by a variety of immigration courts or immigration examiners. The first US Naturalization Test was designed and administered in 1991. But there were many problems with the test. For example, examiners used different sentences for reading and writing, different content for listening and speaking, different levels of difficulty in the sentence writing, unknown passing criteria, and the history and government questions encouraged memorization of discrete facts with little understanding of the material (see Kunnan, 2009, for more details).

Due to these problems, the test was redesigned. The new test, which has been in operation from October 2008, has the following format: (a) applicants would have three chances to read and write a sentence in English based on a vocabulary list, (b) sentences for reading and writing would cover US history and civics, (c) applicants' answers to questions normally asked about their application (N-400) during the interview would form the speaking test. The redesigned test was expected to address the concerns raised with the old test. But so far, all indications are that the redesigned test is no better than the old test as it still requires much memory of facts and figures, very little English language proficiency, and shows continued variability in test items between examiners. This high-stakes test continues to remain a hurdle in the naturalization process for applicants whose first language is not English.

German and Estonian Tests

Germany and Estonia are two countries that have recently enacted legislation that offers citizenship to its residents. In the case of Germany, ethnic Germans (*Aussiedler*) from the

former Soviet Union and other Eastern European countries are granted the right to resettle in the Federal Republic of Germany. But since 1996, prospective *Aussiedler* and their family members have been subject to a German language test that is administered by an immigration examiner. In Estonia, after breaking away from the Soviet Union, the government wrote laws to make the Estonian language the sole official language and required basic Estonian language for Estonian citizenship. The impact of this policy is that a significant percentage of the Russian-speaking residents of Estonia (particularly, the older generations) are likely to be disenfranchised due to the Estonian language requirement. Thus, such high-stakes tests are used by governments to implement and enforce their immigration and citizenship policy.

Major Criticisms and Possible Safeguards

In general, large-scale high-stakes university, college, and workplace (and some immigration and citizenship) tests are often criticized on many grounds especially when these tests impact different groups of test takers, but many safeguards that can eliminate or reduce the disadvantages of high-stakes testing are available. First, one of the sharpest criticisms of high-stakes tests, that they are used for decision making without any additional information, can be blunted if the various stakeholders (score users like admissions officers, employers, citizenship, and immigration examiners) make decisions using many different pieces of information (such as course grades, recommendation letters, job interviews, etc.) and do not rely only on scores from high-stakes tests.

Second, stakeholders (mainly test developers and score users such as admission officers or employers) ought to conduct studies to show that the tests not only have the appropriate qualities but that the consequences of the tests' scores and decisions made based on them are beneficial to test takers, the community, and society in general. If tests are having a detrimental effect on the test-taking community, then more scrutiny of the test would be necessary. In other words, the welfare of the community should be one of the main considerations in using such tests.

Third, necessary systematic validation research studies ought to be conducted by the test development agencies to show to the test users (test takers, test score users, and other consumers of test information) that the tests, their score interpretations, decisions, and related claims can be defended. In fact, these tests need to meet the highest standards of test development, test score interpretation, and decision making. This will go a long way to demonstrate to the stakeholders and the public that the tests are in fact appropriate measures and therefore decisions based on the test scores will be useful and beneficial to the test-using community (see Tindal & Haladyna, 2002, for a discussion of related issues).

Fourth, comprehensive test evaluations need to be conducted on a regular basis for the particular purpose for which the tests are being used. These evaluations need to be based on established standards such as the APA, AERA, and NCME (1999), and include deliberative forums (with policy makers, test developers, test takers, and community members) as recommended by Heubert and Hauser (1999).

Fifth, an independent oversight body that has the authority and responsibility to review tests, test reports, research studies that defend the claims of tests, and overall usefulness of tests could ensure that tests are beneficial to the community (see Madaus, Haney, Newton, & Kreitzer, 1997, for more on this topic).

Sixth, test takers who are unable to take part in high-stakes tests ought to be given alternative assessments so that their career and life chances are not diminished because of the nature of the tests they are required to take. This obviously applies to test takers with disabilities but also to test takers who are expected to take such tests in their second or third languages or in unfamiliar test conditions.

Finally, government (often encouraged by private agencies) need to avoid legislative mandates that violate professional testing and evaluation standards so that new tests that are developed and launched are not in conflict with established ethical and professional practices.

Conclusion

Although high-stakes tests might be useful in many contexts, stakeholders need to understand the limitations of these tests and use multiple sources of information in making decisions on test takers. In addition, the safeguards presented in the previous section regarding conducting research studies and complying with current testing standards are both critical steps in addressing the many criticisms of high-stakes tests. The safeguards are critical as high-stakes tests are here to stay and we have to learn to use them in the best way we can. As Cizek (2005) puts it, "High-stakes tests: we don't know how to live with them, we can't seem to live without them" (p. 50).

SEE ALSO: Assessment of Listening; Assessment of Reading; Assessment of Speaking; Language Testing and Immigration

References

American Psychological Association (APA), American Education Research Association (AERA), National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

Cizek, G. (2005). High-stakes testing: Contexts, characteristics, critiques, and consequences. In R. Phelps (Ed.), *Defending standardized testing* (pp. 23–54). Mahwah, NJ: Erlbaum.

Heubert, J., & Hauser, R. (1999). *High-stakes: Testing for tracking, promotion and graduation*. Washington, DC: National Academies Press.

Kunnan, A. J. (2009). The U.S. Naturalization Test. *Language Assessment Quarterly, 6,* 89–97.

Madaus, G., Haney, W., Newton, K., & Kreitzer, A. (1997). *A proposal to reconstitute the National Committee on Testing and Public Policy for an independent, monitoring agency for educational testing*. Boston, MA: Center for the Study of Testing, Evaluation and Educational Policy.

Tindal, G., & Haladyna, T. (Eds.). (2002). *Large-scale assessment programs for all students*. Mahwah, NJ: Erlbaum.