Teaching to High-stakes Testing in Second Language Learning

HYUN JIN CHO

Purdue University

JUNG HAN

Purdue University

ABSTRACT

The purpose of this review paper is to address high-stakes second language assessment and challenges of teaching to the high-stakes standardized tests. The consequential effect of these large-scale assessments can be both beneficial and harmful in secondary and post-secondary education. Research has shown that teaching to the high-stakes assessment would likely interfere with students' learning in the long run and make students misinterpret their learning ability and academic self-concept with incorrect information from the test results. If teachers focus on teaching to the high-stakes test items, students may not learn the fundamental skills and knowledge that they need to be successful in higher education. Alternatives to the item teaching instruction should be considered in order to provide constructive instructions that students need in their second language learning in the long run.

Keywords: high-stakes assessment, secondary and post-secondary education, second language learning

The fundamental job of a teacher is to promote students' learning. To measure students' improvement in learning, teachers use various types of assessment. Many assessments are meant to assess students' basic skills and knowledge. Further, assessments can measure how students apply their knowledge across different contexts. Standardized tests are designed to assess

students' learning outcomes and skills that they are intended to acquire (Chatterji, 2003; Popham, 2001, 2014; Stiggins, 2002; Volante, 2004). These assessments enable teachers to diagnose students' strengths and weaknesses (McMilan, 2000). Moreover, these tests may help teachers to evaluate whether students reach a certain standard to pass a certain course (Linn, 2000). In addition to strengthening the standard in formal education, standardized tests are often used to compare students and even schools (Kim, 2010).

Most countries employ high-stakes standardized tests to measure students' knowledge and the effectiveness of instruction. However, this type of academic assessment is found to have interfered with the ideal teaching goals (Choi, 2008; Volante, 2004). When the high-stakes assessments are crucial to the students' future, both teachers and students are likely to adjust their classroom activities in response to the tests (Bailey, 1996). Since the results from these assessments are used to label the schools, teachers are more likely to apply test preparation strategies in their instructions. They may use practices on test content or similar items from the high-stakes standardized test.

This influence of the test on the classroom is generally defined as washback effect which can be either beneficial or harmful (Bailey, 1996; Buck, 1988). These tests are often used to compare across students, schools or to determine students' future. High-stakes second language assessment could be one of the examples that show negative washback effect on students' learning and teachers' instructions. The current paper addresses the types of high-stakes second language assessment and alternatives to teaching to the high-stakes assessment.

High-Stakes English Assessments

There are different types of high-stakes second language assessment. There has been an increasing number of students throughout the world who want to study abroad for university in

English countries. The most widely used English language proficiency tests to assess language ability and eligibility of those students are Test of English as a Foreign Language (TOEFL; Test of English as a Foreign Language, 2018) and the International English Language Testing System (IELTS; The International English Language Testing System, 2017). In English-speaking countries, universities depend on tests such as the TOEFL or the IELTS to determine the language ability of students who apply for admission. These tests measure students' language performance against a norm to position test-takers on a continuum across a range of scores. In addition, the tests are usually a standards-based test that use systematic procedures for administration and scoring of large-scale tests (Brown & Abeywickrama, 2010).

In this section, we discussed the two most commonly used English proficiency tests in English as a Second Language (ESL) contexts and the high-stakes English standardized test in Korea in English as a Foreign Language (EFL) contexts.

Test of English as a Foreign Language (TOEFL). The TOEFL measures the ability of non-native speakers of English to use and understand English in academic settings (TOEFL, 2018). The first TOEFL was administered in 1964, and Educational Testing Service (ETS) became responsible for TOEFL, including test development, operation, and finances, in 1973. In 1998, a computer-based version of TOEFL (CBT) was introduced, and administered until September 2006 (Alderson, 2009). Currently, the TOEFL test is provided in two testing formats: Internet-based testing (iBT) and paper-based testing (PBT). TOEFL iBT is the latest version of the TOEFL. This version consists of a reading section taken from university textbooks, a listening section with classroom discussion and conversation, a speaking section with six tasks, and a writing section with two tasks (Alderson, 2009).

The main theoretical basis of the TOEFL iBT is communicative language competence, which is the ability to use the language in complex contexts to complete a variety communication tasks (Alderson, 2009). In line with this theoretical construct, the new version of TOEFL reduced the grammar section and introduced a speaking section and a longer written section to provide more authentic context (Alderson, 2009). As a result, unlike the previous version of the TOEFL which focused on grammar items, the TOEFL preparation classes started to focus on teaching speaking ability and writing skills rather than teaching grammar and vocabulary (Wall & Horak, 2006).

International English Language Testing System (IELTS). Another commonly used test of English language proficiency is the International English Language Testing System (IELTS). The IELTS measures the language ability of candidates who need to study or work where English is used for communication (IELTS, 2017). The IELTS consists of four modules: Listening, Reading, Writing, and Speaking. All test-takers take the same listening and speaking modules, but there is a choice of Academic or General Training Reading and Writing modules. If a test-taker intends to enter undergraduate or postgraduate courses, they are advised to take the Academic modules. If a candidate intends to continue their secondary education in English, to undertake work experience or training, or to emigrate, they are normally advised to take the General Training modules.

The IELTS is internationally focused in the content (IELTS, 2017). Various native-speaker accents such as North American, Australian, New Zealand, and British are used in the listening test, and all standard varieties of English are accepted in test takers' written and spoken responses. Test scores are reported in the form of 'bands' with nine defined levels from non-user (band score 1) to expert user (band score 9) based on overall performance (IELTS, 2017). The

overall band score between 6.0 to 7.0 is considered to be evidence of English language proficiency for university admission, and the minimum is typically an overall score of 6.0 or equivalent (Dooey & Oliver, 2002). Since IELTS' scores have been intended mainly for use in the UK and Australia, research is needed to investigate the appropriateness of scores gained from IELTS as measures of academic language use in North America to ascertain what the test scores mean and how they should be used (Chalhoub-Deville, & Turner, 2000).

Washback of these standardized second language exams can be both beneficial and harmful. It has been suggested that these standardized tests are often used for gate-keeping purposes, and as a result may lead students to emphasize simply achieving an acceptable score rather than developing sufficient language skills (Brown & Abeywickrama, 2010; Chapelle, Enright, & Jamieson, 2011).

Moreover, these English proficiency tests do not necessarily predict students' academic performance. Research has explored the relationship between English proficiency and students' performance (Berman & Cheng, 2010; Feast, 2002; Senyshyn, Warford, & Zhan, 2000; Stoynoff, 1997; Ramburuth & McCormick, 2001). In English proficiency assessments, predictive validity is important in that they are expected to submit evidence of language proficiency. However, the findings produced inconsistent results. In the case of TOEFL, some studies found that TOEFL scores were related to their academic achievement measured by a GPA (Stoynoff, 1997). Students with higher TOEFL scores experienced fewer adjustment difficulties, had more positive experiences, and felt more satisfied than those with lower scores (Senyshyn, Warford, & Zhan, 2000). Also, strong writing skills were correlated with high academic results (Ramburuth & McCormick, 2001). On the other hand, another study found that language proficiency was not significantly related to students' academic achievement (Berman & Cheng, 2010).

These dissimilar findings also appear in the literature on the impact of IELTS scores on performance at university. Feast (2002) found that there was a significant and positive relationship between English proficiency and students' performance whereas other studies found a weak but positive relationship between IELTS and academic achievement (e.g., Kerstjens & Nery, 2000). Dooey & Oliver (2002) mentioned that there was little evidence that IELTS can predict students' performance and academic success. However, inconsistent results from various research studies do not necessarily mean that the assessments are not valid in terms of measuring of English proficiency. Rather, it indicates that language ability can be interpreted as a just one of the contributing factors that predict students' academic success (Feast, 2002).

College Scholastic Ability Test (CSAT). The English subtest in the CSAT is the most high-stakes second language assessment in Korea. This standard test consists of all multiple-choice questions in all subjects. Korea Institute for Curriculum and Evaluation (KICE) develops the test, which is designed to measure the students' academic ability required for college education, commissioned by the Ministry of Education, Science and Technology (MEST). The purpose of the test is to improve fairness and objectivity of student selection by measuring learning abilities and achievements required for college education. Since this test is likely to decide students' college entrance, it has been considered the most high-stakes test in Korean education (Kim, 2010). Since the CSAT determines students' college entrance it also brings up undesirable characteristics and social issues. Many smart test-takers rely on test-taking strategies to receive high scores in the test (Choi, 2008). As a result, some students only focus on improving reading and listening skills compared to speaking and writing skills. Also, this test has been criticized for causing extreme competition among high school students who want to go to universities. Compared to the Scholastic Aptitude Test (SAT) in the United States, the CSAT is

administered only once in a year in November by KICE (Kim, 2010). As a result, many students attend private English institutions in order to achieve high scores on the CSAT English test. Furthermore, another reason to provoke severe competition among students lies in the fact that most Korean universities resort to the CSAT scores to select new students even though other criteria such as interviews, writing essays, and GPAs in high school can be considered (Bae, 2004).

Since the CSAT has been a national high-stakes standard test in Korea, it has influenced the classroom instruction and activities in response to the test. In the case of English education, the high-stakes standard test makes a huge impact on teachers' second language instruction, especially in high schools. The 7th National Curriculum deals with advanced English communication skills, promoting students' four skills in English language learning. Even though the national curriculum in English subjects aims at developing communication skills as an international language, the CSAT English has been focused mainly on listening and reading skills. It has been criticized for lack of items in CSAT English test to assess students' speaking and writing skills compared to other two receptive skills. As a result, English classes in high schools have been focused on receptive skills such as listening and reading to have students succeed in the CSAT, overlooking productive skills such as speaking and writing skills (Kim, 2010). Moreover, the conventional school tests also have been influenced by the CSAT English test. The content of the CSAT English test have a substantial influence on the content of second language classroom instruction (Choi, 2008).

Common Issues of Teaching to High-Stakes Assessments

It is essential for teachers to be aware of the information on the format and structure of high-stakes standard tests so that they can help students get familiar with the test. However,

researchers argue that if teachers spend long time to train students on how to answer particular test items, it is inappropriate (Popham, 2001; Volante, 2004). The researchers pointed out a couple of common issues that occur when teachers apply test preparation strategies in their instruction. They are summarized below.

First, teaching to the high-stakes test can skew the inferences that teachers can make from students' scores and undermines the validity the results (McMilan, 2000; Popham, 2001). Test results should offer teachers and students useful information about students' strengths and weakness. However, if teachers teach to the test, it can inflate students' exam scores, so the result from the tests may not reflect students' actual learning achievement. Whether test preparation is appropriate depends on how much time the teachers should spend and what kind of activities students are asked to engage in (Volante, 2004). If students spend lots of time practicing for the test rather applying the new skills and knowledge, the high scores in the test do not reflect their actual learning. Moreover, students may not be able to acquire the fundamental knowledge and skills that the test does not include. Thus, the test scores do not help teachers or students make valid inferences about students' authentic knowledge or skills (Popham, 2001).

Second, teaching to the test may reduce the depth of instruction in specific subjects and narrows the curriculum, which prohibits students from learning other important components or skills that they might need to in the future (Volante, 2004). Because high-stakes assessment has an influence on curriculum, content, and classroom activities, these areas are likely to be adapted and changed in the direction of the assessment (Amengual, 2010). Students who are good at testing may lack the basic skills and knowledge that are needed to be successful in higher education (Neil, 2003).

Third, teaching to the test may cause negative impact on the teaching profession (Volante, 2004). Teachers may feel pressure to teach the test and give feelings of frustration. Instruction focusing on standardized tests may lead to lack of satisfaction within the teaching profession. On the contrary, language teachers need proper guidelines on the objectives, frameworks, and the implementation of effective pedagogy so that they could feel a stronger sense of self-efficacy in teaching (Malakolunthu & Hoon, 2010).

Fourth, high-stakes assessments give pressure of extra work to prepare to students and their parents. Because these tests are high-stakes and mandatory for secondary students, parents tend to spend money on preparation for these assessments and, thus, feel an extra financial burden (Choi, 2008; Dawson, 2010). Private tutoring that has been employed in preparation for high-stakes assessments provides a clue to reflect on the inadequacies of formal education (Dawson, 2010). Expensive private tutoring has been seen as a response to lower school quality and overregulated formal schooling (Kim, 2004). Not only that, but the negative washback from high-stakes assessment is perpetuated in secondary education (Choi, 2008; Kim, 2010).

Additionally, high-stakes assessments seem to influence students' academic self-concept. Self-concept is considered highly crucial in that it is closely related to students' learning behaviors, academic achievement and self-esteem (Marsch & Martin, 2010). Academic self-concept influenced students' subsequent performance (Shavelson & Bolus, 1982). Inflated high scores may give students false sense of their learning ability (Volante, 2004). Considering the power of high-stakes assessment, the results of the assessment would be associated with the development of students' academic self-concept.

In sum, teaching to the high-stakes assessment interferes with students' learning in the long run and leads students to misinterpret their learning ability and academic self-concept with

incorrect information from the test results. If teachers focus on teaching to high-stakes test items, students may not learn the fundamental skills and knowledge that they need to be successful in higher education.

Desirable Alternatives to Teaching to the Test Instruction

Researchers suggested some desirable alternatives to teaching to the test instruction for teachers. According to Popham (2001), there are two strategies that teachers can employ for high-stakes tests; item-teaching instruction and curriculum-teaching instruction. Item-teaching instruction means that teachers provide their lesson based on the actual items or similar test items whereas curriculum-teaching refers to instruction that is designed based on a specific content knowledge or cognitive skills covered by a given test (Popham, 2001). He suggested that teachers need to receive training in curriculum-teaching which requires them to give their instruction for the content knowledge or cognitive skills rather than teaching test items. When teachers are well aware of the clear description of the curricular content of the high-stakes assessments, they can design instructional activities to promote the knowledge and skills that are required by the assessment (Popham, 2001). In this way, they can lead students to the in-depth discussions for high-stakes tests rather than to test-item instructions (Volante, 2004). Repeated practice or teaching to the test instruction can only increase high-stakes test scores without increasing learners' actual achievement (Shepard, 1990).

Moreover, curriculum-teaching can promote positive washback effects. If teachers provide instruction based on specific content knowledge or cognitive skills that are covered by a test, students will be likely to apply their knowledge and skills in more authentic situations (Volante, 2004). In his review on washback effect, Bailey (1996) summarized some suggestions on promoting beneficial washback. Teachers, administrators, curriculum designers should be

aware of the purposes of the test. The more clearly informative the score of assessment are given to students and teachers, the greater effect the assessment will produce positive washback. Furthermore, it is essential that students think the results are believable and fair. If the assessment is not viewed as relevant to students, its result is unlikely to produce desirable washback to students and teachers. Finally, when teachers use authentic tasks and texts, the test will yield positive washback. For example, if students in English language classroom learn to how to do the authentic tasks, they will be more motivated to learn the language and use it in real life.

Furthermore, teachers can encourage students to perceive the high-stakes assessment as a valuable learning experience. Because the standardized tests are not concerned with students' motivation, it is not certain about whether these high-stakes tests produce the desired improvement to students' learning (Stiggins, 1999). However, recent research demonstrated that there is a positive relationship between students' adaptive beliefs about assessment and their academic achievement (Brown, 2011; Brown & Hirschfeld, 2008; Hirschfeld & Brown, 2009). Positive perspectives about assessments were associated with higher grades and a more positive approach to learning (e.g., Brown & Hirschfeld, 2008). If teachers highlight the constructive role of assessment, high-stakes second language assessments will be perceived as a valuable learning opportunity for students to establish various learning strategies and advance their learning (Cho, 2017). Therefore, creating an environment where students view assessment as a genuine learning process is essential in the long term.

CONCLUSION

The ultimate job of teachers is to promote students' authentic learning so that they can apply the knowledge and skills in many different life situations. The purpose of the standardized

tests is to assess students' achievement in the learning process and measure the effectiveness of instruction. However, as the importance of high scores in the high-stakes standard test has been emphasized, the test results have brought up unexpected negative consequences to students' academic life and even the teaching profession. Research shows ample evidence that teaching to the test items harm teachers' instruction and students' learning as well. More alternatives to the item-teaching instruction should be considered in order to provide constructive instructions that students need in their second language learning in the long run.

ABOUT THE AUTHORS

Hyun Jin Cho is a post-doctoral research associate at the Center for Instructional Excellence at Purdue University. She has a Ph.D. degree in Educational Psychology. She is interested in how students view classroom assessments and what strategies they use for successful academic experiences. She is working on the research regarding motivation in second language environments, beliefs about assessments, and perceptions of their learning environment from a self-determination perspective.

Jung Han is a Ph.D. student in Purdue Polytechnic Institute. She has a master's degree in Literacy and Language Education. Her research interest is language minority students' academic language learning and STEM literacy development.

Inquiries can be directed to Hyun Jin Cho at cho193@purdue.edu

or Jung Han at han336@purdue.edu

REFERENCES

Alderson, J. C. (2009). Test review: Test of English as a Foreign Language™: Internet-based Test (*TOEFL* iBT®). *Language Testing, 26(4),* 621-631. doi:10.1177/0265532209346371

Amengual, M. (2010). Exploring the washback effects of a high-stakes English test on the teaching of English in Spanish upper secondary schools. Revista Alicantina de Estudios Ingleses, 23, 149- 170. doi:10.14198/raei.2010.23.09

Bae, H. S. (2004). The current issues in the Korean scholastic Aptitude Test (CSAT). *Daehakgyoyuk*, *128*, 6-19.

Bailey, K. M. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing*, *13*(3), 257-279. coi: 10.1177/026553229601300303

Berman, R., & Cheng, L. (2010). English academic language skills: Perceived difficulties by undergraduate and graduate students, and their academic achievement. *Canadian Journal of Applied Linguistics/Revue canadienne de linguistique appliquée*, *4*(1), 25-40.

Brown, G. T. (2011). Self-regulation of assessment beliefs and attitudes: a review of the Students' Conceptions of Assessment inventory. *Educational Psychology*, *31*(6), 731-748. doi: 0.1080/01443410.2011.599836

Brown, G. T., & Hirschfeld, H. F. (2008). Students' conceptions of assessment: Links to outcomes, *Assessment in Education: Principles, Policy & Practice, 15*(1), 3-17. doi: 10.1080/09695940701876003

Brown, H. D. & Abeywickrama, P. (2010). *Language assessment: Principles and classroom practices*. Pearson Education. doi:10.1177/0265532207086784

Buck, G. (1988). Testing listening comprehension in Japanese university entrance examinations. *JALT Journal*, *10*, 15-42.

Chalhoub-Deville, M., & Turner, C. E. (2000). What to look for in ESL admission tests: Cambridge certificate exams, IELTS, and TOEFL. *System*, *28*(4), 523-539. doi: 10.1016/S0346-251X(00)00036-1

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2011). *Building a validity argument for the Test of English as a Foreign Language*. New York, NY: Routledge.

Chatterji, M. (2003). *Designing and using tools for educational assessment*. Boston: Allyn and Bacon.

Cho, H. J. (2017). *Promoting International Students' Academic Adjustment* (Unpublished doctoral dissertation). Purdue University, West Lafayette, U.S.A.

Choi, I. C. (2008). The impact of EFL testing on EFL education in Korea, *Language Testing, 25(1),* 36-62. doi:10.1177/0265532207083744

Dawson, W. (2010). Private tutoring and mass schooling in East Asia: Reflections of inequality in Japan, South Korea, and Cambodia. *Asia Pacific Education Review, 11*(1), 14-24. doi:10.1007/s12564-009-9058-4

Dooey, P., & Oliver, R. (2002). An investigation into the predictive validity of the IELTS test as an indicator of future academic success. *Prospect, 17 (1),* 1-19.

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

Hirschfeld, G. H., & Brown, G. T. (2009). Students' conceptions of assessment: Factorial and structural invariance of the SCoA across sex, age, and ethnicity. *European Journal of Psychological Assessment, 25*(1), 30-38.

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

International English Language Testing System [IELTS]. (2017). *The IELTS handbook*. Cambridge: University of Cambridge Local Examinations Syndicate, The British Council, IDP Australia.

Kerstjens, M., & Nery, C. (2000). Predictive validity in the IELTS test: A Study of the Relationship between IELTS scores and students' subsequent academic performance. *English Language Testing System Research Reports, 3,* 85-108. Retrieved September 1, 2018 from https://search.informit.com.au/documentSummary;dn=905790128753688;res=IELHSS

Korea Institute for Curriculum and Evaluation. (2018). *College scholastic ability test*. Retrieved September 3, 2018 from http://www.kice.re.kr/sub/info.do?m=0205&s=english

Kim, T. (2004). Shadow education: School quality and demand for private tutoring in Korea. *KDI School of Pub Policy & Management Paper*, (04-21). Retrieved September 5, 2018 from https://ssrn.com/abstract=635864 or http://dx.doi.org/10.2139/ssrn.635864

Kim, T. Y. (2010). Socio-political influences on EFL motivation and attitudes: Comparative surveys of Korean high school students. *Asia Pacific Educ. Rev, 11*, 211-222. doi:10.1007/s12564-010-9071-7

Linn, R. (2000). Assessment and accountability. *Educational Researcher, 29*(2), 4-17. doi: 10.3102/0013189X029002004

Malakolunthu, S., & Hoon, S. K. (2010). Teacher perspectives of school-based assessment in a secondary school in Kuala Lumpur. *Procedia-Social and Behavioral Sciences, 9,* 1170-1176. doi:10.1016/j.sbspro.2010.12.302

Marsh, H. W., & Martin, A. J. (2011). Academic self-concept and academic achievement: relations and causal ordering. *British Journal of Educational Psychology, 81,* 59-77. doi:10.1348/000709910X503501

McMillan, J. H. (2000). Fundamental assessment principles for teachers and school administrators. *Practical Assessment, Research & Evaluation*, *7*(8). Retrieved October 8, 2018 from http://pareonline.net/getvn.asp?v=7&n=8.

Neil, M. (2003). High stakes, high risk: The dangerous consequences of high-stakes testing. *American School Board Journal*, *190*(2), 18-21.

Popham, W. J. (2001). Teaching to the Test? *Educational Leadership*, *58*(6), 16-21.

Popham, W. J. (2014). *Classroom assessment: What teachers need to know (7th edition)*. Boston: Pearson/Allyn and Bacon.

Ramburuth, P., & McCormick, J. (2001). Learning diversity in higher education: A comparative study of Asian international and Australian students. *Higher education*, *42*(3), 333-350. doi: 10.1023/A:1017982716482

Senyshyn, R. M., Warford, M. K., & Zhan, J. (2000). Issues of adjustment to higher education: International students' perspectives. *International Education*, *30*(1), 17-35.

Shavelson, R. J., & Bolus, R. (1982). Self-concept: The interplay of theory and methods. *Journal of Educational Psychology*, *74*, 3-17. doi: 10.1037/0022-0663.74.1.3

Shepard, L. A. (1990). Inflated test score gains: Is the problem old norms or teaching the test?. *Educational Measurement: Issues and Practice*, *9*(3), 15-22. doi: 10.1111/j.1745-3992.1990.tb00374.x

Stiggins, R. J. (1999). Assessment, student confidence, and school success. *Phi Delta Kappan*, *81*(3), 191-198. Retrieved from October 17, 2018 from http://www.jstor.org.ezproxy.lib.purdue.edu/stable/20439619

Stiggins, R. J. (2002). Assessment Crisis: The Absence of Assessment for Learning. *Phi Delta Kappan*, *83*(10), 758-765. doi: 10.1177/003172170208301010

Stoynoff, S. (1997). Factors associated with international students' academic achievement. *Journal of Instructional Psychology, 24*(1), 56.

Test of English as a Foreign Language (2018). TOEFL iBT Test. Retrieved October 1, 2018 from https://www.ets.org/toefl/ibt/about

Volante, L. (2004). Teaching to the test: What every educator and policy-maker should know. *Canadian Journal of Educational Administration and Policy, 35.* Retrieved October 1, 2018 from http://umanitoba.ca/publications/cjeap/articles/volante.html.

Wall, D., & Horák, T. (2006). The impact of changes in the TOEFL examination on teaching and learning in Central and Eastern Europe: Phase 1, The baseline study. *ETS Research Report Series*, *2006*(1), i-199. doi: 10.1002/j.2333-8504.2006.tb02024.x