The Asian Journal of Applied Linguistics
Vol. 4 No. 1, 2017, pp. 117-128

AJAL

Chinese and Korean college students' perceptions of standardized English tests

Minhee Eom
Department of Writing and Language Studies, University of Texas Rio Grande Valley, USA

Yong Lang
Department of Writing and Language Studies, University of Texas Rio Grande Valley, USA

Caihong Xie
Department of Foreign Languages, Hengyang Normal University, China

This study investigated college students' perceptions of various standardized English tests used in China and South Korea. A total of 357 university students, 195 from China and 162 from South Korea, participated in a survey designed for this study. The survey asked the participants first to select one English test they knew well and subsequently to evaluate its attributes of: test quality, resources, cost, difficulty, test constructs, other skills, and other knowledge. The five English tests most selected were the TOEFL and the TOEIC in Korea and the NCEE, the CET-4 and the TEM-4 in China. Multivariate analyses of variance (MANOVA) and follow-up univariate analyses of variance (ANOVAs) found significant differences between the five English tests in all the test attribute variables except for cost. The comparisons between five English tests provided new insights into significant attributes of standardized English tests used in these two EFL countries.

Key words: English as a foreign language; EFL testing; China; Korea; test attributes

Introduction

Validity is an important concern for all stakeholders of testing practices across all the phases of test development and uses (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AREA, APA & NCME],1999; Bachman, 1990; Messick, 1995). Test-takers are often considered to be passive participants in the process, but their perceptions of the value of a test are an essential part of test development and endorsement (AERA, APA, & NCME, 1999; Brown, 1993). Their feedback and responses are often used to inform definitions and designs of test structures such as testing tasks and scoring rubrics (Attali & Powers, 2010). However, their perceptions of test quality and test attributes have not been much researched even though they are key stakeholders with a direct impact on the success of testing practices.

This study investigates and compares test-takers' perceptions of the values and qualities of standardized English tests used in the People's Republic of China (hereafter, China) and South Korea (hereafter, Korea). In English as a Foreign Language (EFL) countries like China and Korea, the personal and social consequences of testing for individual test-takers have mounted precipitously as their societies continue to place ever-increasing value on English education and English proficiency (Hu & McKay,

ISSN 2308-6262
http://caes.hku.hk/ajal

2012). In such a situation, test-takers' voices cannot be ignored because they are active stakeholders in the testing dynamics of their societies.

Context of the study

This study examines test attributes that were chosen by considering evidence-based arguments on various aspects of testing processes related to construct validity. Since the end of the 1970s, validity has become a unitary concept reinforced by the scientific tradition of validation and has been considered as a pillar of any testing theory, research, and practice (AERA, APA, & NCME, 1999; Bachman & Palmer, 1996, 2010; Chapelle, 2012; Fulcher & Davidson, 2007; Kane, 2012). In contemporary testing theories, the question of validity is a matter of ongoing argument in efforts to support or refute proposed interpretations and uses of test scores in every phase of the testing process. Moreover, the social consequences of testing have become crucial to debates on validity, as the uses and results of tests impact on society and individuals. Any proposed interpretation of test scores must be explicitly positioned within a network of inferences to justify the intended conclusions and any decisions based on the conclusions. Thus, test attributes such as construct representations, internal structures, available resources, and documentation are important factors in the use of tests.

In many EFL contexts, English language education and English testing have serious impacts on the society as a whole as well as on its individual members, a situation that is conceivably due to the rapid globalization that has made English a language of global communication (Hu & McKay, 2012). In countries where English language proficiency plays a critical role in social advancement, including academic and professional promotion, the testing of English language has become an indispensable practice for proving levels of English proficiency.

China and Korea share many similarities in terms of the social impacts and educational practices surrounding English language education (Clavel, 2014). They employ a number of measurement instruments due to an ever-increasing demand for English proficiency. This study examines the variety of English tests available in the two countries and compares their users' perceptions of the quality of those tests.

Testing culture in China and Korea

The English language education systems of China and Korea share quite similar developmental, representational, and social characteristics. China started compulsory English education in the late 1970s when the country opened its doors to the outside world. Subsequent economic reforms and success have brought about rapid and drastic changes and supported a boom in English learning. In Korea, English was a vital part of the post–Korean War reconstruction and the economic developments of the 1970s. In both countries, English is, in practice, a compulsory subject in the college entrance examination and part of coursework for all majors in higher education. Outside academia, English proficiency is often required for those seeking employment or promotion in many fields including education, scientific research, medicine, finance, and business as well as in government or government-supported institutions. As a by-product of this English fever, standardized English testing has become a core element of English education.

Testing has a long history in both countries. In China, it can be traced back to the imperial period of the Han Dynasty (206 BC–220 AD) when tests or exams were developed to select the best candidates to serve as administrative officials in the

government (Cheng, 2008). It could be argued that modern testing and standardized testing originated in China. Korea has a similar history of using tests to choose candidates for government positions. In the Choseun Dynasty (1392–1897), a test called the Kwagoe Shiheom was used for the selection of high-ranking government officials (Choi, 2008). Their long histories of high stakes testing may contribute to the popular acceptance of English testing in both countries.

Another parallel between China and Korea is in the social consequences of English testing. As English is often explicitly or implicitly associated with prestige and social success, English test results play an essential role in giving individuals a step up on the social ladder. They often affect university admission, graduation, opportunities to study or work abroad, obtaining good jobs, career promotion and advancement (Hadid, 2014, October 18; Lee, 2014).

Despite the similarities in the historical development and social impact of English education in China and Korea, there are noteworthy differences in their English testing systems. Compared with that in Korea, English testing in China is more centralized. In China, the National Education Examinations Authority of the People's Republic of China is part of the Ministry of Education and is the exclusive administrative authority on educational examinations at both national and state levels (Hu & McKay, 2012). In consequence, the widely used English tests in China are either national or state tests used for college admissions, college English placement and classes, and the assessment of general communication ability in English. In contrast, English testing in Korea relies heavily on internationally developed EFL tests. Except for the Korean Scholarly Aptitude Test (KSAT), the EFL tests prevalently used in Korea are mostly from international test developers, such as the Test of English as a Foreign Language (TOEFL), and the Test of English for International Communication (TOEIC). There are a few domestically developed English tests in Korea such as Seoul National University's Test of English Proficiency (TEPS), but they are not used as widely as the international EFL tests.

Nevertheless, the EFL tests used in both countries are standardized tests. In China, a special committee consisting of nationally renowned testing experts oversees the designing, development, and modification of tests, and checks the feasibility, reliability, and validity of tests. China also has an English language corpus that was designed and developed to serve as an important source for the country's standardized tests.

Major attributes of language testing

In the field of language assessment, the concept of validity is a core of testing theories and practices. Messick (1989) proposed a unitary concept of construct validity employing the nomological network of construct and observables. Validity arguments should present and integrate evidence and rationales pertaining to particular score interpretations and uses. More recently, Kane (2006) enhanced the argumentative aspect of validity proposing interpretive argument focusing on the score interpretation. Chapelle (2012) pointed out that Kane also emphasized a systematic examination of validity evidence pertaining to the score inference instead of listing types of validity evidence. Advancing the theoretical framework of Messick's validity, Kane's interpretative argument approach shared the familiar goal of defining validity as an appropriateness of test score interpretations and uses.

Various factors and attributes contribute to a validity argument in the process of test development and test uses. A theoretically-defined construct is a core factor of testing (AERA, APA, & NCME, 1999; Bachman, 1990; Bachman & Palmer, 1996, 2010).

However, a language construct is not an easy entity to define precisely. Bachman (1990) distinguished language ability from language skills and argued that these two terms should not be used interchangeable. Language skills are the contextualized realization of the ability to use language in the performance of specific language use tasks. That is, skills are a specific combination of language ability and task characteristics. But language skills like listening, reading, speaking and writing are often used to describe types of ability assessed by a proficiency test.

Construct-irrelevant variables (CIV) can be a serious threat to validity resulting in inadequate score interpretations and test uses (Bachman, 1990; Fulcher & Davidson, 2007; Messick, 1989). Those factors are the variables not directly related to language ability such as individual background knowledge, personality, test-taking skills, and world knowledge systematically, not randomly, affecting the test performances (Bachman, 1990). Test takers must be given sufficient instructions on how to answer the types of questions in the test and overall structures of the test, but if they are getting high scores due to test taking shrewdness, it is a threat to the validity. In addition, test score pollution is a construct-irrelevant factor as it refers to the practice of coaching to boost test scores (Gipps, 1994).

On the other hand, the AERA, APA, & NCME (1999) specify test takers' right and responsibility to prepare adequately for the test, and test developers should provide sufficient resources including guidelines, preparation materials, and test-taking instructions. However, test preparations should not be so excessive as to induce construct-irrelevant effects. That is, it is important to draw a reasonable boundary distinguishing necessary test preparations from excessive coaching leading to test score pollution.

The provision of proper resources can be considered an issue of fairness. Educational Testing Service (ETS) Guidelines for Fairness Review of Assessments (ETS, 2009) require that construct-irrelevant personal characteristics of test takers have no significant effect on test performance or their interpretation. The fairness guidelines ensure providing impartial access to products and services and impartial registration, administration, and reporting of assessment results. Those guidelines also list socioeconomic status as a factor to be considered in ensuring fairness validity along with other social factors such as gender, ethnicity, and more (ETS, 2009). The findings of previous research suggest that a socio-economic factor influences students' academic achievements when they attend schools with a plethora of educational resources and support (Ethington & Wilson, 2010).

In sum, validity is a key theoretical concept backing up the overall quality of language testing, and all stakeholders including test-takers are to be involved in the on-going process of validation. This study focused on those attributes about which test takers could express opinions through their experience of preparing for or taking the English tests available in their educational setting.

Aim of the study

This study investigated college students' perceptions of standardized EFL tests used in China and Korea in regard to testing attributes of: test quality, resources, cost, difficulty, test constructs, other skills, and other knowledge. Despite the similarities in historical and social practices of English education cultures in China and Korea, the types of English tests used in the two countries are considerably different. Because of the differences in test purposes, constructions, and structures, it is impossible to make direct comparisons of the English tests used in two countries. However, the comparative

analysis of perceptions of the test takers may reveal the relative strengths or weaknesses of those tests especially in relation to the test features, and/or their validity considerations.

Methodology

Participants

A total of 357 university students (195 from China and 162 from Korea) participated in the survey. The Chinese participants were second year undergraduates majoring in English in a large public university located in the southern-central part of the country. The Korean students were English major undergraduates from a university located in a major southern city in Korea.

Instrument

Data were collected via a survey developed by the authors and approved by the Institutional Review Board at the researchers' institution. The survey had three sections: Part one collected demographic information about the participants; part two investigated participants' perceptions of various aspects of the English tests available in their country; and part three examined participants' perceptions of the individual and social consequences of English testing in their country. This paper is based on the data derived from the first two parts of the survey.

Part two of the survey was further divided into two sections. In the first section, the participants were asked to rate their familiarity and experience with various tests. Based on literature reviews and personal conversations with teachers and students, the authors had created a list of EFL tests that the participants could be expected to be familiar with or to have studied for. The participants were asked to rate the tests on a scale from 1 to 6 (1 = I don't know it, 2 = I know it a little, 3 = I know it, 4 = I know it well, 5 = I have studied for it, 6 = I have taken it), the higher the numeric value, the greater the familiarity with the test. The participants who rated any test with a value of 4 or higher were asked to move onto the second section of part two to answer questions about their perceptions of the testing attributes. Their responses in this section were the main data for this study. The reliability Alpha of part two of the survey was .778.

Based on the preliminary analysis of familiarity of various English tests, a total of five English tests were selected. The two English tests receiving most responses in Korea were the Test of English as a Foreign Language (TOEFL) and the Test of English for International Communication (TOEIC). Three tests were selected among those used in China. The first, receiving most responses, was the National College Entrance Examination (NCEE), which is used for university admissions and one of the most frequently used English tests in China (Graddol, 2013; Wang, 2006). The test receiving the second most responses was the Band 4 (CET-4) component test of the College English Test (CET), which is a mandatory English test for non-English majors including three component (Zheng & Cheng, 2008). The third highest rated test was TEM-4 for freshmen and sophomores, a component of the Test for English Majors (TEM) which is a criterion-referenced English language test specifically targeted at undergraduates majoring in English language and literature (Jin & Fan, 2011).

Variables: Seven test attributes

In order to investigate the perception of testing attributes, the survey questions were classified into seven test attributes.

Test quality (TQL)

The general perception of test quality is an important attribute as it is related to the concept of validity. The survey contained the following four related items:

- I think it is a reliable test.
- I would recommend it to others.
- It is a well-recognized test in our society.
- The test developer is trustworthy.

Cost (CST)

This variable was included to examine whether financial factors can be a distinguishing factor of EFL testing. The following two survey items were used:

- The test fee is reasonable.
- The test fee is expensive.

Resources (RSC)

The availability of test preparation materials was included as a variable to examine differences in the English testing contexts of the two countries. The following two items asked about this variable:

- It is easy to acquire test prep materials.
- Sufficient test preparation courses are available.

Test difficulty (DIFF)

Two items were used to examine the overall difficulty of a test.

- The test is too difficult.
- The test is easy.

Test constructs (CON)

Due to the two countries' different procedures of test development and standardization, students in China and Korea may have different perceptions of how well the English tests of their country evaluate language skills and knowledge. Because of the complexity of such test constructs, more survey items examined this variable than any of the others:

- I think this test properly evaluates academic English ability.
- I think this test properly evaluates reading ability.
- I think this test properly evaluates grammar knowledge.
- I think this test properly evaluates vocabulary knowledge.
- I think this test properly evaluates speaking ability.
- I think this test properly evaluates writing ability.
- I think this test properly evaluates listening ability.

Other skills (OSK)

Many non-language factors can be involved in test preparation and test-taking practices. This variable examined the perceived involvement of test-taking skills in English testing with three items:

- Good time management skills improve the score.
- Good test-taking skills improve the score.

- The more often I take the test, the higher score I get.

Other knowledge (OKNG)
This variable used two items to represent construct-irrelevant factors specifically related to non-language knowledge:

- Other topical/subject knowledge improves the score.
- Other cognitive skills, like IQ, are helpful to improve the score.

Data analysis methods
Multivariate one way analyses of variance (MANOVA) were used to identify significant differences between the seven test attributes across five English tests from the two countries (TOEIC, TOEFL, CET-4, NCEE, and TEMP-4). The follow-up univariate Analysis of Variance (ANOVA) were also conducted to identify the pair that showed differences.

Results
As seen in Table 1, the *resources* variable received the highest total ratings (M = 4.57, SD = .984) with the highest rating in the TOEIC (M = 5.05, SD = .816) in Korea followed by the NCEE in China (M = 4.81, SD = .873). The *test quality* variable received the second highest total ratings (M = 4.41, SD = .789), where the TOEFL test received the highest rating (M = 4.80, SD = .697) followed by the NCEE (M = 4.65, SD = .821).

Table 1. Descriptive statistics of ratings of test attributes

			TQL	CST	RSC	DIFF	CON	OSK	OKNG
China	CET-4 (n = 57)	Mean	4.15	3.68	4.66	3.22	3.67	3.64	4.03
		SD	.735	.552	.844	.643	.713	.539	.783
	NCEE (n = 96)	Mean	4.65	3.81	4.81	3.15	3.68	3.83	4.35
		SD	.821	.601	.873	.638	.809	.581	.873
	TEM-4 (n = 188)	Mean	4.25	3.79	4.21	3.30	3.86	3.84	4.16
		SD	.766	.577	.995	.557	.757	.502	.718
Korea	TOEFL (n = 17)	Mean	4.80	3.97	4.59	3.53	4.74	4.31	3.94
		SD	.697	.819	.972	.695	.814	.670	.568
	TOEIC (N = 106)	Mean	4.49	3.68	5.05	3.42	3.84	4.08	3.55
		SD	.740	.607	.816	.465	.699	.685	.794
Totals		Mean	4.41	3.77	4.57	3.31	3.86	3.89	4.02
		SD	.789	.593	.984	.575	.777	.610	.833

Note: TQL = test quality; CST = cost; RSC = resources; DIFF = test difficulty; CON = test constructs; OSK = other skills; OKNG = other knowledge.

In the Levene's test for the homogeneity of variance, two variables were found violating the assumption of homogeneity of variance: *difficulty*, F (3, 457) = 2.77, p = .027, and *other skills*, F (4, 457) = 2.60, p = .035. Thus, Pillai's trace was chosen as the best multivariate measure, and the following univariate analyses used corrected post hoc tests (as described by Leech, Barrett, & Morgan, 2014).

The multivariate analysis revealed a significant difference among the EFL tests, Pillai's trace = .478, F (454, 1816) = 8.81, p = .000, η = .35. The effect size of the finding (η = .35) was large or larger than typical (using the guidelines of Cohen, 1988), so we can infer that the students' perceptions of the EFL tests in the two countries were significantly different in the attributes we investigated. In order to identify significantly different pairs of tests on an attribute, the follow-up univariate ANOVAs were conducted with the Games-Howell post hoc test because the homogeneity assumption was violated.

Table 2. Results of the Games-Howell post hoc tests

	Test (A)	Test (B)	Mean Dif. (A-B)	Std. Error	p
Test quality (TQL)	NCEE	CET-4	.479	.130	.003**
		TEM-4	.392	.102	.001**
	TOEFL	CET-4	.636	.195	.023*
		TEM-4	.549	.178	.042*
Resources (RSC)	TEM-4	CET-4	-.470	.131	.004**
		NCEE	-.597	.116	.000***
		TOEIC	-.838	.108	.000***
Constructs (CON)	TOEFL	CET-4	1.077	.219	.000***
		NCEE	1.064	.214	.000***
		TEM-4	.878	.205	.003**
		TOEFL	.904	.209	.003**
Other skills (OSK)	CET-4	TOIEC	-.435	.098	.000***
		TOEFL	-.669	.178	.008**
	TOEIC	CET-4	.435	.098	.000***
		NCEE	.257	.089	.034*
		TEM-4	.233	.076	.021*
Other knowledge (OKNG)	TOEIC	CET-4	-.482	.129	.003**
		NCEE	-.798	.118	.000***
		TEM-4	-.615	.093	.000***

Note: * p < .05 ** p < .01 ***p < .001

In the univariate ANOVA, five out of seven variables showed a statistically significant difference: *test quality* (TQL), F (4, 460) = 7.49, p = .000, η = .241; *difficulty* (DIFF), F (4, 460) = 3.67, p = .006, η = .176; *test constructs* (CON), F (4, 459) = 7.93, p = .000, η = .257; *resources* (RSC), F (4, 460) = 16.27, p = .000, η = .354; *other skills* (OSK), F (4, 459) = 8.25, p = .000, η = .263; and *other knowledge* (OKNG), F (4, 457) = 15.71, p = .000, η = .348. In term of the effect size (η), it was large or larger than typical for *resources*, *other skills*, and *other knowledge*, medium or typical for *test quality* and *test constructs*, and small or smaller than typical for *difficulty*. The *cost*

variable was not found significantly different between the tests, F (4, 460) = 1.48, p = .207. The post hoc analysis results reported in Table 2 include only the variables that were statistically different with typical and larger than typical effect sizes.

The post hoc analyses showed that the TOEFL received the highest rating of all on *test quality*, with significant differences from the CET-4 and TEM-4. Among the three tests used in China, the NCEE was rated significantly higher on *test quality* than the others. Also, In terms of *resources*, the TEM-4 was significantly different from the other Chinese tests and the TOEIC. This finding indicates that study resources are considered relatively less available for the TEM-4 than for the other tests. The TOEIC was significantly different from all of the Chinese EFL tests in the *other knowledge* variable. The mean differences indicate that the Chinese English tests might involve significantly more non-language knowledge such as subject knowledge than the TOEIC. Regarding *test constructs*, the post hoc tests revealed that the TOEFL was perceived by the participants as significantly different from (i.e., better than) the rest of the tests. As for the *other skills* variable, the TOEIC showed a significantly higher mean than all the Chinese tests suggesting that test-taking skills are perceived as much more important for the TOEIC than for the Chinese English tests.

Discussion

This study examined five EFL tests widely used in China and Korea to see how the participants perceived the variables of test quality, cost, resources, test difficulty, test constructs, other skills, and other knowledge. Except for *other knowledge*, the ratings given by the Korean students were significantly higher than those given by the Chinese students. Overall, this finding indicates that the test-takers in Korea have more positive perceptions of the English tests they take. The tests used in Korea are for the most part internationally well-accepted, standardized English proficiency tests designed by leading international test developers. This finding could raise a flag for Chinese test developers and researchers about the need for more research on critical features of their domestically developed English tests to enhance valid uses of the tests.

Other knowledge was the only variable that received significantly higher ratings from Chinese participants. This finding implies that the Chinese participants have a stronger belief that non-English knowledge such as subject area knowledge and content knowledge can help them achieve higher test scores. In particular, the post hoc results revealed that all the three Chinese English tests showed significantly higher ratings than TOEIC on the *other knowledge* variable. These differences mean that Chinese English tests might involve significantly more non-language constructs than the TOEIC. Given the fact that the CET-4 and TEM-4 are based on the national curricula for English courses, they are more like achievement tests that assess subject areas of English studies than English proficiency tests according to common classifications of tests (see, among others, Bachman, 1990; Bachman & Palmer, 1996). Strictly speaking, the English tests used in China may not be English proficiency tests, which could be problematic if they are used for the purpose of assessing English proficiency. The findings of this study, therefore, suggest that the appropriateness of these tests for proficiency assessment should be re-evaluated, and that ways of strengthening their validity should be considered.

The fact that the *other skills* variable received significantly higher ratings in Korea indicates the perceived importance in Korea of test-taking skills such as computer skills and cognitive strategies. One possible explanation is that Chinese English tests are paper-based, and test-takers may feel that test-taking strategies are more important in

taking computer-based tests. However, several research studies have found no meaningful effects of computer familiarity on test performance (Choi, Kim, & Boo, 2003; Taylor, Jamieson, Eignor, & Kirsch, 1998), and speculation on a computer-effect is beyond the scope of this study. Nonetheless, it is alarming if test-takers perceive that skills other than English proficiency significantly affect their test results. The post hoc analysis results show the TOEIC has significantly higher ratings on this variable than all the other English tests in China. As a matter of fact, test preparation is a big business in Korea. One TV commercial for a TOEIC preparation school even describes the TOEIC not as an English test but as a skill (Youngdanki, 2013). This advertisement, which has run on national TV, is an example of how the media contribute to seriously misleading public perceptions of the test. In Korea, the TOEIC is a symbol of the importance of test-taking strategies and drills, rather than authentic communication ability. The findings of this study confirm the problematic perceptions of the TOEIC in Korean society. Test-takers have rights to obtain the information about "test-taking strategies including time management, and advisability of omitting an item response", but the use of test-taking skills should be for obtaining valid responses (AERA, APA, & NCME, 1999, p. 85). Thus, this finding calls for further investigation on the proper use of TOEIC in Korea.

As for the *test constructs* variable, the post hoc tests revealed that the TOEFL was perceived as a better test than any of the other tests included in this study. Despite the small size (n = 17), this finding may demonstrate a notable pre-eminence of the TOEFL as a test of English language. As many studies have reported (e.g. Chalhoub-Deville & Deville, 2005; Chalhoub-Deville & Turner, 2000), the Educational Testing Service is a globally well-known test developer that conducts serious validity research on their tests. The findings of this study confirm that, in Korea, the TOEFL is considered the gold standard of EFL tests.

Similarly, findings in regard to the *test quality* variable showed that the TOEFL was perceived as better than the CET-4 and TEM-4. Interestingly, the Chinese NCEE was also rated significantly higher on *test quality* than the CET-4 and the TEM-4, indicating that the participants considered the NCEE more trustworthy than the other tests.

Finally, the TEM-4 received significantly lower ratings than the other English tests on *resources*. This suggests that far fewer study resources are available for the TEM-4, which may be due to the limited test-taker population because the test is only for English majors. More resources are developed for the more widely used tests, whether for profit or more efficient centralized management. However, when a test is in use, general information including helpful materials should be provided to test takers (AERA, APA, & NCME, 1999) suggesting further availability of resources for TEM-4 would be desirable to protect the rights of test takers as well as to enhance its valid uses.

Conclusion

Despite the similarities in the social and educational impact of English testing in China and Korea, the types of standardized English tests available in the two countries are quite different restricting direct comparisons of these tests. Nonetheless, the analyses of students' perceptions allow us to compare certain features of the standardized English tests used. The findings from this investigation indicate that the English tests used in China are perceived differently from the TOEIC and TOEFL used in Korea. The differences might be due to the different purposes and constructs of the tests, but questions still arise about the validity of the Chinese tests. It can be argued that the China-developed English tests like the NCEE, the CET and the TEM are not English

proficiency tests but more like achievement tests assessing content knowledge of English language. This study also calls for further considerations on such aspects as construct representations and resources for those tests. In addition, when compared with other tests, the TOEIC seems to engage more test taking skills than the other tests. Thus, further research is desirable for the proper use of TOEIC in Korea.

This study did not include domestically developed English tests used in Korea as there was insufficient data. It would also be useful in future research to make direct and within-group differences in the perception of the testing attributes, and to interview participants for richer understanding of their perceptions. In spite of these limitations, the present study reveals some special features of standardized English tests used in China and Korea, and provides some new insights into valid uses of those tests.

About the authors

Minhee Eom is an associate professor of Applied Linguistics in the Department of Writing & Language Studies at the University of Texas Rio Grande Valley. She teaches undergraduate and graduate courses in descriptive linguistics, language and culture, assessment, and research methodology. Her research interests include language assessment, ESL/EFL education, quantitative research designs and third language acquisition.

Yong Lang is a full professor of Applied Linguistics in in the Department of Writing & Language Studies at the University of Texas Rio Grande Valley. He teaches undergraduate and graduate courses in descriptive linguistics, language and culture, fundamentals of language development, and sociolinguistics. His research areas include linguistics, applied linguistics, second language acquisition, and cross cultural communication.

Caihong Xie is an associate professor of English in the College of Foreign Languages at Hengyang Normal University. She teaches undergraduate courses in translation and business English. Her research interests and publications are mainly in the areas of translation theories and practice.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington: AERA Publications Sales.

Attali, Y., & Powers, D. (2010). Immediate feedback and opportunity to revise answers to open-ended questions. *Educational and Psychological Measurement, 70*(1), 22-35.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford: Oxford University Press.

Brown, A. (1993). The role of test-taker feedback in the test development process: Test-takers' reactions to a tape-mediated test of proficiency in spoken Japanese. *Language Testing, 10*(3), 277-301.

Chalhoub-Deville, M., & Deville, C. (2005). A look back at and forward to what language testers measure. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 815-832). Lawrence Erlbaum: Mahwah, N.J.

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.

Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple…. *Language Testing, 29*(1), 19-27.

Cheng, L. (2008). The key to success: English language testing in China. *Language Testing, 25*(1), 15-37.

Choi, I.-C. (2008). The impact of EFL testing on EFL education in Korea. *Language Testing, 25*(1), 39-62.

Choi, I.-C., Kim, K. S., & Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing, 20*(3), 295-320.

Clavel, T. (2014, January 19). China, South Korea face familiar woes in English quest. *The Japan Times*. Retrieved from http://www.japantimes.co.jp/community/2014/01/19/issues/china-south-korea-face-familiar-woes-in-english-quest/#.VgxKfvlViko

Educational Testing Service. (2009). *ETS guidelines for fairness review of assessments*. Princeton, USA.

Ethington, C. A., & Wilson, T. (2010). Mathematics achievement and African-American students in urban schools. *Investigations in Mathematics Learning, 2*(2), 19-32.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. London and New York: Routledge.

Gipps, G. V. (1994). *Beyond testing: Towards a theory of educational assessment*. New York: Routledge Falmer.

Graddol, D. (2013). *Profiling English in China: The pearl river delta*. Cambridge: Cambridge English Language Assessment.

Hadid, A. (2014, October 18). English education in Korea: Unrealistic expectations. from http://thediplomat.com/2014/10/english-education-in-korea-unrealistic-expectations/

Hu, G., & McKay, S. L. (2012). English language education in East Asia: Some recent developments. *Journal of Multilingual and Multicultural Development, 33*(4), 345-362. doi: 10.1080/01434632.2012.661434

Jin, Y., & Fan, J. (2011). Test for English majors (TEM) in China. *Language Testing, 28*(4), 589-596.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport, CT: Praeger Publishers.

Kane, M. (2012). Validating score interpretations and uses. *Language Testing, 29*(1), 3-17.

Lee, C. (2014). TOEIC adds to stress for young job seekers. from http://www.koreaherald.com/view.php?ud=20140326000917

Leech, N. L., Barrett, K. C., & Morgan, G. A. (2014). *IBM SPSS for intermediate statistics: Use and interpretation* (5 ed.). London & New York: Routledge.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education/Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

Taylor, C., Jamieson, J., Eignor, D., & Kirsch, I. (1998). *The relationship between computer familiarity and performance on computer-based TOEFL test tasks. ETS Research Report (RR-98-08, TOEFL-RR-61)*. Princeton, New Jersey: Educational Testing Service, USA.

Wang, X. B. (2006). An introduction to the system and culture of the college entrance examination of China. *College Board Research Notes, RN-28*.

Youngdanki. (2013). TOEIC is a skill from https://www.youtube.com/watch?v=XWlKR_54OXg

Zheng, Y., & Cheng, L. (2008). Test review: College English Test (CET) in China. *Language Testing, 25*(3), 408-417.