Revisiting Test Validation Models: 
Implications for the Reform of the National Testing of Japanese University Applicants

Izumi WATANABE-KIM
International Christian University
Sophia University

Abstract

With the anticipated 2020 reform of national testing of university applicants, test validation will play an even greater role in the Japanese educational context. English language assessment, in particular, has received more attention due to the inclusion of an oral communication test. In such times of change, this paper aims to: 1) provide a brief historical overview of how the concept, issues, and practice of test validation have evolved in the last century, and 2) offer suggestions as to how Japanese universities might proceed in selecting from an array of commercially available English language tests for admission purposes. Beginning with Messick’s (1989) definition of validation, this paper will touch upon key concepts of both old and new validation models, including criterion-model, content-model, construct-model, and the argument-based validation model. Research findings call for a careful, systematic, and ethical approach towards fulfilling the demands of the upcoming changes that should take into consideration fairness to test takers, as well as potential washback of test use.

Keywords: Japanese university entrance examination, test validation model, argument-based approach to validation, fair use of test, washback

Recently, the National Center for University Entrance Examinations announced that the testing of Japanese university applicants will undergo a major reform to be completed by 2020. The English language portion of admissions testing has received much attention due to the anticipated inclusion of an assessment of oral communication. In light of this reform, the Japan Language Testing Association (JLTA) submitted a proposal in 2016 to the Ministry of Education, Culture, Sports, Science and Technology (MEXT) to promote transparency, fair use, equal opportunity and positive washback in English language teaching practices (JLTA, 2016). This may reflect anxiety not only amongst testing specialists but also among the common stakeholders, especially test takers, whose futures may be determined by a high-stakes test. Historically, testing has been associated with the potential for discrimination and other forms of social injustice. To avoid these pitfalls among others, it may be appropriate to revisit the concept of test validity. This paper thus aims to provide a brief
historical overview of test validation in the United States. First, Messick’s (1989) definition of validation will be presented. Next, this paper will review the three most documented approaches to test validation: criterion-model, content-model, and construct-model, following the taxonomy provided by Messick (1989) and Kane (2006). The paper will conclude with a discussion of Kane’s (1992, 2006, 2013) argument-based approach to validation, and will offer six suggestions as to how it can be applied in the context of the upcoming reform.

**Historical Overview**

In his seminal paper, Messick (1989) defined test validity as “the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores” (p. 13). He stressed that it is the “interpretation” and the “use” of test scores which must be questioned, not the test itself, though the process of validation should not overlook the “trustworthiness” of the instrument. Although the validity model evolved from the criterion to the unified model of construct validity, the underlying issue for the last century remained the same. This was the question of whether or not the test was measuring what it intended to measure, and if so, to what degree.

**The Criterion-Model of Validation**

At the beginning of the 20th century, a marked development in mass testing and scoring methods was observed in the United States (Shepard, 2016). Starting with the formation of the College Entrance Examination Board in 1900 (Zwick, 2006), Binet’s development of the IQ test in 1905, and Thorndike’s launching of the first standardized achievement test in 1908, all signified remarkable advancements in psychological and educational testing (Shepard, 2016). Such instruments were enthusiastically promoted as they served societal needs in a rapidly industrializing world (Shepard, 2016). The push to measure mental aptitude was largely influenced by eugenicists who aimed to match one’s mental capacity with an appropriate role (i.e., occupation) in society (Shipman 1994). For example, the Army Alpha test, a descendant of the IQ test, was developed to serve the demands of military screening during World Wars I and II (Zwick, 2006).

A similar movement was also seen in the field of education during the first decades of the 20th century when tests were used to sort students into “highly differentiated curricula” (Shepard, 2016, p. 113) to cope with the influx of immigrants in the United States. The idea was to manage and allocate educational resources by sorting out those with a low mental aptitude -- who would only require vocational training -- from those who tested as gifted and would then be placed on the fast-track to college (Shepard, 2016; Zwick, 2006). Unfortunately, and not surprisingly, the mass-testing movement led to social injustice; testing masked racial discrimination by playing the role of gatekeeper when it came to social opportunities and advancement (Shipman, 1994). The early concept of test validation did not include the notion of fairness and washback of test use.

In those days, test validation was (and to some extent, continues to be) about finding
the correlation between test scores and external variables or criterion. For instance, if an employer seeks someone who would be a good match for a particular job, a criterion is identified for that job. For a typing job, for instance, the criterion may include some observable measures such as speed and accuracy of typing. Next, a test task requiring typing skills would be designed, and validation of the test would then involve determining whether the actual job of typing performance agreed with the test results (Cureton, 1951).

Similarly, a set of test scores and student grades could be compared. If a correlation was found, the test in question might be a valid measure of predicting the academic success of students, as seen in the current practice of college admissions in the U.S. wherein scores on standardized tests (e.g., SAT and ACT) are examined in relation to students’ high school grade point average (GPA). In fact, most validation studies in the first half of the 20th century were “predictive in nature” prior to the introduction of the content-validity in the 1950s (Lissitz & Samuelsen, 2007, p. 438).

The Content-Model of Validation

The fundamental issue that the criterion-model faced was the validity of the criterion itself, which is supposed to be the basis for validation (Kane, 2001, 2006, 2008). To address this issue, Cureton introduced a new concept, known as content-validity, in 1951 (Lissitz & Samuelsen, 2007). The content-model “interprets test scores based on a sample of performances in some area of activity as an estimate of an overall level of skill in that activity” (Kane, 2006, p. 19). In other words, it examines how well the test samples the domain of knowledge, skill or activity that is intended to be measured. For example, a history test may cover a particular target domain, a period in history (e.g., the Renaissance) or a historical event (e.g., Pearl Harbor).

Similarly, in a language proficiency test, a representative sample of target language use (TLU), such as the use of the past-progressive form, may be tested by means of a storytelling task. Achieving validity with this model requires sufficient evidence that the target domain is well-represented in the test content. Although it is possible to question validity by claiming that the sample is biased in some way (Kane, 2006), the content-model generally works well for testing specific knowledge and skills as it does not require external criteria (whose validity itself may be questionable) for validation (Cronbach, 1971; Kane, 1982, as cited in Kane, 2006). To date, the content-model is employed most commonly in the validation of achievement tests.

Test development under the content-model begins with specifying the target language use (TLU) domain, which serves as a blueprint for the test. The blueprint must describe what falls within the province of knowledge/topics or skill/task of the target domain. As seen in the study of Revesz’s (2012) which was analyzed in Purpura, Brown, and Schoonen (2015), a language test designed to measure the degree of mastery of the past-progressive form must draw out a list of language forms of the TLU domain. The items in the actual test are then compared with the test specifications drawn from that list (Purpura et al., 2015).
tasks may be designed to elicit the use of a particular domain, such as the past-progressive. In her study, Revesz (2012) prepared a ‘grammaticality judgment task,’ a ‘written picture description task,’ and ‘oral description tasks’ to assess the mastery of the past-progressive form.

As one might expect, the content-model has its limitations. Because it does not involve external measures in the process of validation, the content-model is often criticized for being subjective (Kane, 2006; Messick, 1989). For example, when test developers are conducting the validation, they may be more likely to claim an agreement between the TLU domain and the testing instrument because the item writers may be reluctant to criticize their creations (Kane, 2006). Even when an external expert makes a judgment, the issue remains: how does one overcome subjective evaluation of a representative sampling? Messick (1989), who later proposed the unified-model of validation, states that content-validity has its value but is limited in scope, and must be used in combination with other methods to achieve a higher degree of validity.

The Construct-Model of Validation

The concept of construct validity was first introduced by Cronbach and Meehl (1955) and defined as “some postulated attribute of people, assumed to be reflected in test performance” (p. 283). It was a new approach to measurement, proposed as a means of addressing unobservable psychological traits, such as “anxiety” or “ego,” when neither criterion nor content was available to base the test on (Kane, 2001, 2006). Psychological traits had to, and continue to, be treated more or less as concepts, rather than as observable skills or knowledge to be displayed and measured. Cronbach (1971) defined construct as “a theory that sketches out the presumed nature of that trait” (Cronbach, 1971, p. 462 as cited in Kane, 2006, p.20). Hence, validating the construct means to validate the underlying theory, which explains the “nature of that trait” (Cronbach, 1971, pp. 462–463 as cited in Kane, 2006, p.20). Only when the theory is proven valid can the test be considered as a valid measure of that trait.

As stated, construct validation is an analysis of a theory, which involves finding relevant evidence to prove a particular theory (Kane, 2006). It is perceived as a combination of the criterion-model and the content-model of validation with a theory (construct) to prove. For instance, if one were to measure “anxiety” with this approach, s/he first specifies the construct of “anxiety” and the underlying theory which explains/describes the psychological attribute. Test development proceeds based on that specification of that theory and the validation process involves multiple components (Cronbach, 1971 as cited in Kane, 2006). Finding correlation(s) may only be a part of a larger network of study to support a theory, and the network itself is “interactive” as one part may be influencing (or influenced by) other parts of the network of evidence (Embretson, 2007, p. 452). Other components may include a review of content, statistical analysis for reliability, finding relationships to other external variables and considering alternative interpretations of the theory (Kane, 2006).
By the late 1980s, the construct-model had evolved into a unified model, integrating all three (criteria, content, and construct) models of validation, with construct being the overarching concept. The validation effort is now seen as “developing a scientifically sound validity argument” (AERA et. al., 1999, p. 9 as cited in Kane 2006, p. 23) In other words, it is a unified effort to present a coherent argument supported by multiple components of available evidence to justify the test use and interpretation of test scores.

The Argument-Based Approach to Validation

By the new millennium, Messick’s (1989) unified model of construct validation had undergone another transformation. Though the unified model had integrated a variety of evidence types to support score interpretations, it was lacking in providing clear guidelines as to how the evidence should be analyzed and presented (Kane, 2001). Kane’s (1992) argument-based approach to validation addressed that problem by providing a framework for the validation efforts (Sawaki, 2011). Kane (2006) introduced the idea that “to validate a proposed interpretation or use of test scores is to evaluate the rationale for this interpretation or use” (p. 23). He proposed a two-tier process in forming an “interpretive argument” and a “validity argument” for intended test use. In the first tier, one describes how the test scores are interpreted. The second tier entails explaining why the interpretation is acceptable. A comprehensive picture of the multi-step process of validation is shown in Figure 1 below.

In the left column, the six steps involved in the validation process are listed: conceptualization, evaluation, generalization, extrapolation, utilization, and implication. In each step, arguments are made in a two-tier process (middle and right columns). As mentioned above, an interpretive argument is followed by a validity argument. For example, if one wishes to interpret in the conceptualization stage that the test tasks represent the TLU domain, evidence (such as the results of a pilot test) is gathered to make a validity argument.

<table>
<thead>
<tr>
<th>Validation process</th>
<th>Interpretive argument</th>
<th>Validity argument</th>
</tr>
</thead>
<tbody>
<tr>
<td>6. Implication</td>
<td>The test brings positive washback</td>
<td>Washback study</td>
</tr>
<tr>
<td>5. Utilization</td>
<td>The test can be the basis of decision</td>
<td>Standardized test scores</td>
</tr>
<tr>
<td>4. Extrapolation</td>
<td>The test resembles real life situation</td>
<td>Student samples &amp; follow-up interviews</td>
</tr>
<tr>
<td>3. Generalization</td>
<td>Results are consistent and generalizable</td>
<td>Statistical analysis</td>
</tr>
<tr>
<td>2. Evaluation</td>
<td>Test and scoring procedures are appropriate</td>
<td>Expert opinion &amp; post-test interview</td>
</tr>
<tr>
<td>1. Conceptualization</td>
<td>Tasks represent TLU domain</td>
<td>Pilot test &amp; revision</td>
</tr>
</tbody>
</table>

*Figure 1. The process of argument-based validation (adapted from Chapelle, Enright, & Jamieson, 2008).*
As one would expect, the six steps are linked to one another that they may be presented as a series or chain of arguments. Kane’s (1992) argument-based-validation model is based on the informal argument, a model presented by the philosopher, Stephen Toulmin. As shown in Figure 2 below, Toulmin’s (1958) model of argument follows the “if-then rules” (Kane, 2013, p. 451) which presents a claim as presumptive in nature.

For instance, one might claim that “students are likely to be successful in the English language courses at university” if “they have (already) passed the Eiken level 2 test” since “the level of English courses are comparable to Eiken level pre-1.” The claim is supported by datum (in this case, the passing of Eiken level 2) and a warrant (the level of English courses) with a backing (of document review). The claim is plausible unless there is a counter-argument (that students are not performing well in English courses) with a backing (high rate of failure). To engage in a rebuttal (that failure is not related to language proficiency), further backing (e.g., evidence to support a claim that that failure is actually due to low attendance) is required to make another claim.

In each of the six steps of the argument-based validation, a claim (such as the one above) is made, and the claim is examined against evidence following the logical flow of the Toulmin’s model. Thus, in the first (conceptualization) stage, an argument is formed employing Toulmin’s terminology and logical flow to convince the audience that the test task resembles real-life language use. Following the first claim, a second argument is formed (in the evaluation stage) to convince the interested parties that the test and scoring procedures are appropriate. Subsequently, a third (generalization) argument is formulated to reiterate that the test is reliable. The construction of these arguments is repeated up to the sixth (implication) stage, and in each stage, arguments are formed in the Toulmin’s style. In this way, the efforts of validation at various stages are presented as a series of arguments to convince the audience that the test is a valid instrument to use for a particular purpose. An additional explanation of the application of these models will be presented in the following sections, wherein the models will be applied in the context of selecting an appropriate test for university admission (see Suggestions 1 to 6 for the application of the argument-based validation framework, and Suggestion 4 in particular, for the use of Toulmin’s model of argument).

As expected, test validation efforts will not reach a definitive conclusion regarding their absolute validity. Instead, these efforts will only inform the audience about whether an interpretation is “more-or-less plausible” (Kane, 1992, p. 533). Therefore, the main focus of test validation is on making a convincing argument to justify score interpretations and use of the test (Kane, 2013). Also, it is worth noting that the argument-based validation model may be the first attempt to fully incorporate test implication in the validation process. This reflects how the validation practice has evolved to embrace the notion of fairness to stakeholders, especially, to test takers. Incorporating the concept of fairness to test takers implies that regardless of how accurate or reliable the measurement is if testing does not bring about a desirable social and educational outcome, the test is not “valid.”
Figure 2. Toulmin’s (1958) model of argument (adapted from Sawaki, 2011)

Application of Argument-based Validation

Currently, Kane’s (1992, 2006, 2013) argument-based approach appears to represent the mainstream of test validation. It is more common to see the application of Kane’s framework with large-scale standardized tests, such as the TOEFL.

For instance, when a new listening test was designed for the new TOEFL project led by Chapelle et al. (2008), the team consulted a large corpus of spoken academic English as the basis for creating test materials to increase authenticity (Enright et al., 2008). The corpus included over one million words and some non-native English pronunciation, reflecting the reality of classrooms in the United States. This initiated a debate about whether to include non-native pronunciation so as to increase authenticity, or to exclude this variable in order to maintain fairness to all test takers. (Enright et al., 2008). The assumption was that test takers would perform better on listening tasks when the speaker and the test takers shared the same L1. In other words, if the listening materials included the speech of Japanese speakers of English, Japanese test takers would have an advantage over other test takers with a different L1.

However, when the effect of non-native pronunciation on test takers’ performance was examined, it was reported that “Spanish listeners performed best when the speaker was Spanish”, but surprisingly, “Japanese listeners performed worst when the speaker was Japanese” (Major, Fizmaurice, Bunta, & Balasubramanian, 2002 as cited in Enright et al., 2008, p. 103). This disparity may be attributed to their past experiences listening to English. While Spanish speakers encounter Spanish speakers of English in their daily lives, Japanese
speakers’ exposure to spoken English may be limited to audio materials, which are more likely to be recorded by native speakers of English. As a result, it appears that TOEFL decided to exclude non-native speakers of English from the listening materials to avoid potential bias. This could be interpreted as an effort to maintain fairness to test takers at the expense of authenticity of the listening materials.

Major et al.’s (2002) study was only one of many projects completed in the initial stages of the new TOEFL. Such thoroughness of test validation would not be possible without the resources of the Educational Testing Service (ETS). Most test developers at Japanese universities would resort to writing conversation scripts for a listening test based on their intuition and experience as resources allotted for such projects are limited.

However, not all tests need to undergo such an extensive validation process. For example, if the test is given in a listening class to assess students’ ability to identify /l/ and /r/, a test with minimal pair items may serve the purpose rather than having students listen to an authentic conversation which includes those sounds (Sawaki, 2011). Most will agree that in this case, there is no need to consider ‘extrapolation’ and ‘implication.’ Therefore, prioritizing the components of validation is important, and the purpose of the test and the context shall determine what to prioritize (Sawaki, 2011).

Discussion

In light of the upcoming 2020 policy for the university entrance examination, it is necessary to be aware of recent developments in language testing theory and validation. As universities will select commercially available tests (e.g., Cambridge, Eiken, GTEC, IELTS, TEAP, TOEFL, and TOEIC) for admission, it is critical to monitor the validation process of these tests in order to make informed and responsible choices.

There are two major concerns pertaining to predictive validity. First, there is no index of criteria which specifies the types of English learners skills required for study at the university level. Secondly, the tests mentioned above measure different constructs, which in turn makes it difficult to compare the results of these tests to determine the cut-off score for admission.

Koizumi (2018) presents a practical solution to these issues by listing six questions (see Table 1 below) for decision-makers to consider when selecting a test. She suggests choosing different tests within the same university because the desired linguistic abilities (the criteria for selection) may vary from department to department. For example, an international liberal arts department may want to place more emphasis on oral communication skills, whereas science and engineering departments may want to require higher reading comprehension skills (Koizumi, 2018). She also points out the importance of considering the anticipated washback from the start of the decision-making process.
Table 1
*Questions to Ask when Choosing a Test for Admission (Koizumi, 2018, p. 186-189)*

<table>
<thead>
<tr>
<th>Question Number</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>By employing the test, what type of positive washback is anticipated?</td>
</tr>
<tr>
<td>2.</td>
<td>Aside from the anticipated positive washback, what is the purpose of test use?</td>
</tr>
<tr>
<td>3.</td>
<td>What information (relating to linguistic abilities) do you seek to obtain from the test?</td>
</tr>
<tr>
<td>4.</td>
<td>Which assessment method can make it possible to obtain such information?</td>
</tr>
<tr>
<td>5.</td>
<td>Of the assessment method listed in Q4, is it necessary to use a standardized-test?</td>
</tr>
<tr>
<td>6.</td>
<td>If yes, is the existing test appropriate for the given test use?</td>
</tr>
<tr>
<td>6-1</td>
<td>Is there a test that provides the information we seek?</td>
</tr>
<tr>
<td>6-2</td>
<td>Is it possible to anticipate positive washback by using that test?</td>
</tr>
<tr>
<td>6-3</td>
<td>Is the test appropriate for test use?</td>
</tr>
<tr>
<td>6-4</td>
<td>Can the test fully measure the linguistic abilities of the applicants?</td>
</tr>
<tr>
<td>6-5</td>
<td>Is the test level appropriate for the applicants?</td>
</tr>
<tr>
<td>6-6</td>
<td>Are the test tasks reflecting the target language use domain?</td>
</tr>
<tr>
<td>6-7</td>
<td>Is there enough validation information provided by the test developers?</td>
</tr>
<tr>
<td>6-8</td>
<td>Is it logistically possible to administer the test?</td>
</tr>
<tr>
<td>6-9</td>
<td>Are there any other concerns we must consider?</td>
</tr>
</tbody>
</table>

These questions will be referred to again in the next section following the framework of argument-based validation.

Another major concern relating to employing the commercially available tests for university admission is the content-validity of these tests. Since they are not developed for Japanese university admission, test content does not conform to the guidelines of MEXT as it is now stated with the current National Center Test. (Watanabe, 2013). Testing specialists and high school teachers, in particular, are concerned with the compatibility with the high school curriculum once universities start employing the commercially available tests (Nikkei, 2018).

MEXT, on the other hand, seems to anticipate a positive washback effect on the high school curriculum by incorporating oral communication in the university entrance examination. Considering that the university entrance exam was held accountable for the lack of communicative abilities of Japanese English learners, it may be a natural course of events for MEXT to take drastic measures to start employing a 4-skills English test for university applicants. However, as JLTA (2016) notes, empirical research on washback indicates that changing the test is unlikely to result in a revision of the high school curriculum. A review of the teacher-training programs and measures to increase assessment literacy of in-service teachers may be necessary to initiate such a change (JLTA, 2016).

Perhaps, the most serious concern is related to the construct-validation. University staff involved in the selection of commercially available tests may initially ask “Which English test shall we use?” rather than “What English language abilities do we want to measure in
prospective applicants that will be compatible with our courses?” Without a clear idea of the construct, the decision may be influenced more by practical concerns such as the cost, availability, and popularity of the test, as well as by the choices of competing universities. Certainly, these elements are not to be disregarded, but the primary consideration needs to be the core issue: the test construct.

Test constructs are publicly available and accessible. For example, the construct of the TOEFL oral communication test can be found on the website of ETS and states that “the speaking section measures test takers’ ability to speak English effectively in educational environments, both inside and outside of the classroom” (ETS, 2018). When compared with the description of TOEIC, it is fairly obvious that the two are measuring different language domains. According to TOEIC, their tests are “valid assessments of English-language speaking and writing skills for business. Speaking skills are important for effective presentations, face-to-face communication, meetings, video-conferencing, tele-conferencing and telephone conversations” (ETS, 2018). But this information will not help university admission officers select a test without an understanding of what takes place in the actual university English classroom. As the test is supposed to measure the “readiness” of prospective applicants, it is fundamental that decision makers be aware of what applicants need to be ready for.

Suggestions for Selecting a Test

The following six suggestions are made from the perspective of a Japanese university English language practitioner. They are intended to be compatible with the framework of argument-based validation (see Figure 1).

Suggestion 1: Define the TLU domain

Prior to selecting the test, it is crucial for the university to know the target language use domain. In order to map the type and level of English language skills required for a particular context, it might be useful to conduct a document review such as a review of the syllabi of an English language program and/or observations of classes in order to ascertain their learning objectives. A review of textbooks can also provide valuable information such as the level of English literacy required to be successful in such classrooms.

For example, in the EAP program where the author works, first-year undergraduate students are required to use an in-house textbook which includes a collection of academic texts; these comprise over 100,000 words, wherein the proportion of academic vocabulary is approximately 8.4% (Watanabe-Kim, 2015). For a prospective student to be successful in such a program, how much literacy is required? One way to answer this question may be to find an estimate of the vocabulary size of incoming students. It is said that for readers to have 70% comprehension, they must understand approximately 98% of the vocabulary in a text (Schmitt, Jiang, & Grabe, 2011), and an estimate of the average vocabulary size of Japanese university students is 3,715 word families (McLean, Hogg, & Kramer, 2014).
So which test would provide the most information needed to determine the “readiness” of a student? Such questions (see Q3 and Q4 of Table 1) need to be asked not only to assess reading ability, but for all four English language skills. Having a clear picture of what is required in university English classrooms will help in determining which test is appropriate.

**Suggestion 2: Evaluate the Commercially Available Tests**

Once a sketch of the TLU domain is drawn, it is possible to select candidates of the test to be used for admission purposes. For an EAP program, such as the one mentioned above, tests developed to measure academic English skills such as the TOEFL or IELTS academic module may work best. Because reading passages in these tests are chosen from “introductory college-level” (ETS, 2018) texts used in English-medium universities, the level of reading is comparable to those in the in-house textbook of the program.

After potential tests are shortlisted, the next step would be to examine the scoring procedures and relative weight of test components. Studies about scoring are typically conducted by the test developers and results should be made public. One responsibility of the university staff who is in charge of test selection is to check them to determine if the test is appropriate for the purpose. Questions to ask in this stage are listed under Q6 above (see Table 1).

For example, if one is evaluating the scoring procedure of the written portion of the test, it would be helpful to check the scoring rubric. An expert (e.g., an actual writing instructor) can be called upon to evaluate the scoring rubric to assess whether there is coherence between the rubric and the writing goals of the department, course, or program.

**Suggestion 3: Check how the Test Developers Generalize the Test Results**

As expected, standardized-test developers conduct studies prior to launching their products in order to claim that the test results are consistent and generalizable. Therefore, it is the responsibility of the decision makers (or people in charge of test selection) at a university to check to ensure that the test is a “reliable” instrument to be used for admission purposes. Test developers must be willing to explain the reliability and generalizability of their product to all universities (if not already made available to the public).

In order to better understand how the quality of the tests are controlled, universities can invite test developers for an explanatory meeting. For example, in the EAP program where the author works, an expert from IELTS was invited to the faculty retreat to host a seminar on how to score speech samples of applicants. It was a hands-on experience for the English language instructors to appreciate how the scoring of speech is standardized. Given the high-stakes nature of the test, it is crucial that decision makers and teaching staff at the universities are informed of the reliability and generalizability of these tests. This is an essential prerequisite for test consideration.
Suggestion 4: Check if the Test Resembles Real-Life Situations

Similar to suggestion one, universities will want to ensure that the test resembles real-life situations, in this case, that the TLU domain is drawn from English language classrooms. To compare the test tasks and the actual language use, student samples of language production (in writing or speaking) of prototypical students can be examined against the test tasks. It is also possible to recruit prototypical students to pilot the test and conduct follow-up interviews to incorporate the point of view of test takers in the decision-making process.

For instance, let us consider a hypothetical situation: A university is considering the use of TOEFL for admission purposes. Upon taking a closer look at the test, university staff may find a discrepancy between the oral test task and the reality of English language classrooms. While TOEFL speaking tasks require test takers to process academic lectures and reading materials, real-life oral tasks (in the English language classroom) may only involve responding to basic conversational stimuli such as “Do you agree or not agree with whaling?” In such a case, TOEFL is not an appropriate test to use because the test measures the readiness of learners to study abroad in an English-medium university. TOEFL speaking is a linguistically and cognitively demanding task which may not be appropriate for high school graduates who are yet to develop such information processing skills even in their mother tongue.

To give another example, imagine a situation where most university applicants pass the Eiken level 2, an assumed sign of readiness for the university English courses. If this assumption is, in fact, true, one might make the claim that applicants are likely to be successful if they have already passed the Eiken level 2 because the materials in the English curriculum at this particular university are comparable to Eiken level pre-1. Upon admission, students enroll in English classes, but a large number of students seem to fail those classes. Why? Are they not ready to handle these English courses? When looked at more closely, one might find that the failure is not related to language proficiency; rather, it is related to motivation. Failing students are not attending classes due to their part-time work commitments. In such a case, the original claim, employing the Eiken test for admission, might still be valid. Figure 2 shows the flow of argument described above using the Toulmin (1958) model.

Suggestion 5: Consider if the Test can be the Basis of Decision for Admission

Once the test meets the basic criteria of reliability, generalizability, and authenticity, the next step would be to consider whether the test score(s) can be utilized as a part of the admissions decision-making process. One way is to find out how the test performs is by looking into the distributions of test scores. If the scores are clustered towards the higher/lower end, it may be a sign that the test is not level-appropriate for the target demographics, and, therefore, is not a valid measure to use for decision-making. If it discriminates well, then it can be used not only for admission purposes, but also for
placement into different streams of English language instruction. This can be an important benefit as a considerable amount of time and energy will be saved on the side of both university administration and test takers if placement tests can be waived by using test scores for both admission and streaming.

**Suggestion 6: Consider the Washback of Testing**

Lastly, but most importantly, attention must be paid to washback of test use. As Koizumi (2018) suggests (see Table 1), the implication of test use must be considered from the start of the test selection process.

However, as much as teachers would like to avoid “teaching to the test,” it is also known that teachers modify their instruction to meet the demands of a test. For instance, if the TOEIC test is chosen by a university, it is highly likely for those seeking admission to that university to prepare for the TOEIC, which is designed to measure “English-language skills needed in the workplace” (ETS, 2018). Preparing for a TOEIC speaking task might involve imagining and imitating a conversation between business personnel which could take place in a copy room of an office building. This may be challenging for high school students who have never been in such a situation. The likely solution might be to memorize multiple scenarios so that when a picture of business personnel in a copy room is presented, test takers can recite the memorized script. Clearly, this is not a desirable washback of a speaking test as it does not develop their ability to engage in a spontaneous conversation in a business context.

But if a test is well-designed, as well as level-appropriate and the test construct is in alignment with the TLU domain and the learning objectives of an institution, why not teach to the test? For instance, if a scoring rubric for an oral communication task contains desirable learning outcomes, the rubric can be used to monitor the development of one’s speaking abilities. If a test can bring positive washback to instruction and learning, testing can become a welcome addition to the curriculum.

**Conclusion**

This paper attempts to show how selecting the “right” test for a particular context is closely related to test validation. Though test developers and test users may have different perspectives, both are essentially making the same inquiry. They are seeking to find out whether or not a test score is reflective of the knowledge/ability it intends to measure, and if so, whether it serves the purpose without bringing negative washback. Specifically, this includes: 1) finding coherence between testing and the domain of instruction and/or language use and 2) ensuring/monitoring the quality of the test while 3) taking into consideration the implications of test use.

As shown in the historical review of validation models, the current concept of test validation has evolved to embrace the notion of fairness to test takers. Lessons from the past
have taught us how testing can potentially lead to social injustice and discrimination. Bearing this in mind, many have voiced their concerns that employing commercially available tests for university admission could result in the unfair treatment of applicants. In addition, since testing opportunities are limited in rural areas, and the cost of these tests tend to be high, economic disparities may result in unequal educational opportunities, which in turn may result in further economic inequality. This is certainly beyond the scope of English language practitioners, but we must acknowledge our responsibility to strive for providing equal educational opportunities by selecting an appropriate test(s) for admission. A careful, systematic, and ethical approach to test selection should include voices from the field of English language practitioners, test developers, language testing professionals and finally, the test takers.

Acknowledgements
I would like to express my appreciation to Audrey Morrell, Robert MacIntyre, Yoshinori Watanabe, and two anonymous reviewers for their helpful suggestions on earlier versions of this paper.

References


