11-27-2018

Standardized Language Proficiency Tests in Higher Education

Gary Ockey
Iowa State University, gockey@iastate.edu

Nazlinur Gokturk
Iowa State University

Standardized Language Proficiency Tests in Higher Education

Abstract

In higher education, standardized academic language proficiency test scores are often used for multiple purposes, including admissions of international students to degree programs and identification of students' post-entry language support needs. In this chapter, issues surrounding the use of high-stakes standardized academic language proficiency tests for making decisions about international English as a second language (ESL) students are explored. Specifically, (a) stakeholders' views and knowledge about standardized academic language proficiency tests, (b) predictive validity of standardized academic language proficiency tests for academic success, (c) the use of standardized language test scores for placement into language support courses, and (d) the use of locally developed tests for placement into language support courses rather than standardized tests are discussed. Based on the discussion, suggestions for the appropriate use of standardized language test scores for making admissions and placement decisions are provided.

Keywords

Standardized language tests, Admissions, ESL placement, Higher education

Disciplines

Bilingual, Multilingual, and Multicultural Education | Educational Assessment, Evaluation, and Research | Higher Education | Language and Literacy Education

Comments

This accepted book chapter is published as Ockey G.J., Gokturk N. (2019) Standardized Language Proficiency Tests in Higher Education. In: Gao X. (eds) Second Handbook of English Language Teaching. Springer International Handbooks of Education. Springer, Cham. Doi: 10.1007/978-3-319-58542-0_25-1. Posted with permission.

The Uses of Standardized Academic Language Proficiency Tests in Higher Education

Gary J. Ockey and Nazlinur Gokturk

Iowa State University

Ames, Iowa, USA

Abstract

In higher education, standardized academic language proficiency test scores are often used for multiple purposes, including admissions of international students to degree programs and identification of students' post-entry language support needs. In this chapter, issues surrounding the use of high-stakes standardized academic language proficiency tests for making decisions about international English as-a-second-language (ESL) students are explored. Specifically, a) stakeholders' views and knowledge about standardized academic language proficiency tests, b) predictive validity of standardized academic language proficiency tests for academic success, c) the use of standardized language test scores for placement into language support courses, and d) the use of locally developed tests for placement into language support courses rather than standardized tests are discussed. Based on the discussion, suggestions for the appropriate use of standardized language test scores for making admissions and placement decisions are provided.

Key words: standardized language tests, admissions, ESL placement, higher education

Introduction

Standardized academic language proficiency tests are commonly used to make decisions about test takers in academic settings. Standardized tests are expected to have gone through a rigorous test development process, are based on a set of test specifications, use a justifiable scoring model, and follow universal procedures for administration (Davies, et al., 1999). They are also usually supported by empirical investigations for their stated purpose. Examples of standardized academic language proficiency tests include the Test of English as a Foreign Language internet-based Test (TOEFL iBT: http://www.ets.org), the International English Language Testing System (IELTS: https://www.ielts.org), and the Pearson Test of English Academic (PTE Academic: https://pearsonpte.com).

While it seems reasonable to assume that when standardized academic language proficiency tests are used for their intended purpose, the abilities that are inferred for a particular score are appropriate, such a claim may not always hold true. Because these tests are designed for fairly general contexts, they may not produce accurate scores for all types of test takers and contexts for which they were developed. For instance, students with little experience in completing multiple-choice items may not be able to demonstrate their English ability on a multiple-choice standardized test designed to determine if students have enough English ability to study in an English-medium university. Moreover, English language skills necessary for academic success at undergraduate and graduate levels may vary across disciplines, programs, and/or institutions, and a standardized English language proficiency test may not be appropriate for identifying the appropriate threshold of English needed for all of these contexts.

A further complication in the use of standardized language tests in higher education is that the tests are commonly used for purposes for which they were not primarily designed. When this is the case, interpretations about a test taker's ability are even less likely to be appropriate. An obvious example would be the use of standardized academic language proficiency test scores to determine placement of students into leveled math classes. Most would agree that this would likely lead to inaccurate decisions about the individual's math skills. On the other hand, it would be much less clear as to whether the use of these standardized academic language proficiency test scores could appropriately be used for a more closely associated purpose, such as identifying students' university post-entry language support needs. The two purposes, identifying language needs for placement and language proficiency for university admission, are similar; however, they are not equivalent in that placement testing involves placing students at levels of a pre-existing language program, whereas proficiency testing requires measuring how much ability and/or capability a test taker has in a given language (Cheng and Fox, 2017).

Despite concerns about the use of standardized academic language proficiency tests in higher education, their use for making high stakes decisions is prevalent. In fact, scores from these tests are commonly used for multiple purposes, including ones for which they were not originally developed. The use of standardized academic language proficiency tests may be tempting for test users because they can limit the need for resource intensive locally developed tests. Users commonly rely on standardized language test scores for admissions decisions, a purpose for which they have been designed and may be appropriate, as well as to determine if there is a need for placement into language support courses, a purpose for which they were not designed, and are less likely to be appropriate. Given the potential advantages and concerns about uses of standardized academic language proficiency tests in academic settings, an exploration about what is and is not known about these tests may provide stakeholders guidance for how to use these test scores most effectively for their particular contexts.

In this chapter, some of the issues surrounding the use of high-stakes standardized academic language proficiency tests used in academic settings are explored. The chapter begins with a discussion of stakeholders' views and knowledge about interpretation and use of standardized academic language proficiency test scores in academic settings. Next, the usefulness of these standardized language tests for predicting subsequent academic success is examined. This is followed by a discussion about the use of standardized language test scores to place students into appropriate levels of language support courses. Finally, the use of locally developed tests for ESL placement rather than standardized tests is debated. The discussion is limited to international ESL students, and most claims made in the paper are based on empirical investigations of TOEFL iBT and IELTS because these are the tests that have been most used and studied for these purposes in higher education.

Stakeholders' Views and Knowledge about Standardized Academic Language Proficiency Tests

Broadly speaking, stakeholders can be categorized into two groups: those who make decisions about the use of a particular test and those who are affected by these decisions. Because stakeholders' understanding and uses of test scores may positively or negatively influence test consequences they play a critical role in determining the appropriateness of a test for a particular purpose (Messick, 1996; Bachman and Palmer, 2010). In this section, the focus is on the views of stakeholders who generally make decisions about the uses of standardized language tests, that is, individuals that have the power to decide how standardized language test scores are used in their institutions. To a lesser extent, the views of other stakeholders who may have little or no input on decisions regarding the use of test scores, the test takers themselves, are also discussed.

Several studies have been conducted to investigate stakeholders' interpretations and uses of standardized language tests in higher education contexts (e.g., Baker et al., 2014; Ginther and Elder, 2014; Hyatt and Brooks, 2009; O'Loughlin, 2008, 2013; Rea-Dickins et al., 2007). A recurring finding among these studies is that most decision makers lack sufficient knowledge about standardized language tests to make informed decisions about the use of these tests for admission and placement purposes in their institutions.

O'Loughlin (2011) examined stakeholders' knowledge and beliefs about the use of IELTS for selecting international students and planning students' future language learning in an Australian university. Data were collected from university policy documents, questionnaires administered to 20 administrative and academic staff, and in-depth semi-structured interviews with 10 of these stakeholders. The researchers found that there was little empirical basis for setting entry-level IELTS scores, and that the stakeholders had insufficient knowledge about the meanings of IELTS scores. Moreover, the scores were barely utilized to guide students' future language learning, with the exception that students with an overall IELTS score of 6.0 or less, but admitted into linguistically challenging courses, were required to take two additional ESL courses.

In a much larger study, funded by the Educational Testing Service, similar conclusions about stakeholders' knowledge of test scores were drawn. Ginther and Elder (2014) investigated stakeholders' interpretations and uses of TOEFL iBT, IELTS, and PTE scores in two American and Australian universities. Four-hundred and eighty-one administrative and academic stakeholders completed questionnaires and 30 of them participated in follow-up interviews. The researchers reported that most stakeholders had generally limited knowledge about the meanings of test scores. For instance, there was some confusion among the stakeholders regarding the meaning of minimum cut-scores, the “the line between success and failure” on the test (Davies, et al., 1999), established by their institutions.

Coleman et al. (2003), who contrasted students’, instructors’ and administrators’ views about IELTS in three English-medium universities in Australia, the United Kingdom, and China, also reached similar conclusions. Four hundred twenty-nine students and 195 instructors and administrative staff responded to questionnaires, and 19 students and 18 instructors participated in semi-structured interviews. The findings indicated that in all three institutions instructors and administrative staff, who were likely the ones making decisions about the uses of standardized tests scores, were not very knowledgeable about the test and the meanings of IELTS scores. In addition, the research indicated that the instructors had concerns about the predictive validity of IELTS scores due to factors such as excessive test preparation and the non-disciplinary nature of the test. Unlike the other two studies, this one also explored the views of another group of stakeholders, the test takers, themselves. In contrast to the beliefs of the instructors, it was found that students considered IELTS as an accurate and adequate measure of academic language proficiency.

While some score users are very knowledgeable about the appropriate uses of standardized academic language proficiency test scores, research suggests that most score users are not sufficiently informed about how to use these scores for making decisions. This suggests the need to build language assessment literacy, knowledge and skills that enable an individual to “understand, evaluate, and, in some cases, create language tests and analyze test data” (Pill and Harding, 2013, p. 382), among score users. Stakeholders, including policy makers, instructors, and administrators, need to be assessment literate to make informed decisions about assessment practices in their own contexts (Inbar-Lourie, 2008, 2013; O’Loughlin, 2013; Popham, 2004;

Taylor, 2009, 2013). Given that the lack of assessment literacy among stakeholders may influence not only the validity and ethicality but also the feasibility and effectiveness of the tests, it is of utmost importance for stakeholders in higher education institutions to develop their assessment literacy (Baker, 2016; Taylor, 2009; O'Loughlin, 2013). To ensure valid use of test scores for admission and/or placement purposes, score users must familiarize themselves with standardized language tests and meanings of test scores. In addition, they must recognize the importance of setting cut scores for admission and placement decisions for their specific contexts based on empirical investigations rather than intuitive judgments. It should be kept in mind that even though interpretations made based on test scores may be valid for a particular purpose, cut scores set inappropriately may lead to incorrect decisions, compromising the usefulness of a test for its intended purpose.

While the responsibility for developing language assessment literacy and gathering validity evidence to support decisions made based on test scores rests with local decision makers, it is the test developer's responsibility to provide test users with test manuals that report test development and validation procedures as well as score interpretations on various test taker populations. Test developers should also make research available on issues related to test design that might have an influence on score interpretations. For example, employing a different test format rather than multiple-choice could minimize the potential impact of test format on score interpretations (Bachman and Palmer, 2010).

The research also suggests that decision makers do not generally believe that standardized test scores can be used to effectively guide students' future learning. This seems to be a defensible approach, given that the primary aim of these standardized academic language tests is to help determine the extent to which an individual has sufficient English ability to succeed in an English-medium university. That said, it may be possible to better exploit the results of standardized tests to identify students who might need additional language support. For example, in a local context where curriculum objectives are generally aligned with standardized tests, newly admitted students' test scores might be appropriately used, at least to a certain extent, to guide their future language learning. On the other hand, it could be that standardized tests scores may not be very useful for placing students into levels of a pre-existing ESL course in a particular context, since they are unlikely to be aligned with the objectives of the curriculum in that context.

It is also interesting to note the rather large gap between students' and instructors' perceptions about the usefulness of standardized language tests for placement purposes. While instructors tend to believe that standardized academic language proficiency test scores may not be useful for identifying students who need additional language support, students generally believe that these tests can be used to make placement decisions. For instructors, the unwillingness to use standardized test scores for placement purposes may stem from a perceived mismatch between students' test and real life academic language performance, or a general view that standardized tests are inherently bad. For students, the belief that standardized test scores can be used for placement decisions might stem from their experience with the high stakes nature of these tests; they are commonly used for high stakes decisions, such as admissions. Students might naturally assume that they must be appropriate for other language assessment purposes, given their general importance. This belief is likely further supported, in a student's mind, by the fact that many universities exempt students who obtain high standardized test scores from taking in-house placement tests. In short, teachers may not trust standardized tests for post-entry placement decisions because they are standardized tests, which are not designed to be used for placement purposes, while students may trust standardized tests for placement purposes because standardized tests have likely had major impact on their study habits and academic lives and consequently they have been indoctrinated to trust them.

Predictive Validity of Standardized Academic Language Proficiency Tests

Designed to provide evidence of an L2 test taker's academic language ability, standardized academic language proficiency tests would presumably elicit scores that could be used to predict academic success. Predictive validity refers to "the correlation between test scores and later performance on something" (Carr, 2011), which in the case of standardized language tests would be the correlation between these test scores and measures of academic success, such as course grades. Surprisingly, at least to some, research in this area has generally failed to find a strong relationship between standardized language test scores and college success measures. For example, in their study of 113 undergraduate Business students at an Australian university, Kerstjen and Nery (2000) found only weak correlations between IELTS scores (overall and section) and GPA, with correlations ranging from .12 to .20. In another small-scale study conducted on 65 students in three disciplines, Business, Science, and Engineering, Dooey and Olive (2002) found that, in general, IELTS scores had no significant correlations with GPA. The only exception was IELTS reading scores, which related weakly with course grades in three disciplines, with correlations ranging from .21 to .37. As part of a study on the criterion-related validity, that is, the extent to which scores on one test relate to scores on an accepted indicator of the ability to be assessed (Davies, et al. (1999), of PTE Academic, Riazi (2013) investigated the relationship between 83 students' PTE Academic scores and their first-year GPAs at an Australian university. Reported correlations of PTE scores with grades ranged from .28 to .35, indicating a weak relationship between PTE scores and first-year academic performance.

Ushioda and Harsch (2011) examined the relationship between 95 pre-sessional graduate students' IELTS scores and their course grades. The results indicated a weak to moderate relationship between test scores and grades, with correlations ranging from .26 to .58.

Using a much larger sample than these other studies, Cho and Bridgeman (2012) investigated the predictive validity of TOEFL iBT, also using first-year university GPA as a criterion of academic success. Data included 2,594 students' academic records gathered from 10 American universities. The findings showed that TOEFL iBT scores accounted for only 3% of the variance in GPA (correlations of roughly .09) for the sample, and even when disciplinary variation was taken into consideration, TOEFL iBT scores were found to explain only an additional 3-4% of the differences in GPA. In a follow-up study, Bridgeman et al. (2016) examined the relationship between TOEFL iBT scores and GPA, taking into account linguistic background, academic discipline, and section score profiles. Data collected from 787 undergraduate students in an American university were analyzed. The researchers found that when all students were included in the same analysis, TOEFL iBT total scores correlated only weakly with GPA (r =.18), providing further support that standardized academic language proficiency tests are not good predictors of college success.

A number of possible reasons for the rather low predictive relationship between standardized language test scores and university success have been postulated. For one thing, these two purposes imply two related but somewhat different constructs. Standardized academic language proficiency tests aim to measure general language proficiency, while academic success as measured by GPA or a similar measure is impacted by not only language ability, but a number of other factors, such as motivation, learning strategies, disciplinary knowledge, and academic acculturation (Cheng and Fox, 2008; Fox et al., 2014). In other words, standardized test scores may be predictive of academic success, but these other factors are so important that the correlations between the two constructs are quite small. Another potential explanation for failure of standardized test scores to predict academic success relates to the challenges of measuring academic success. GPA may not be a very good measure of academic success since a great amount of variation in grading practices exists, even within a given program in one university. Moreover, grade inflation may lead to a rather restricted range of grades, which could also limit the power of an analysis to identify a relationship. An additional reason for such results might be due to homogeneous samples of test takers: none of the studies reported here included test takers who did not attend university. This group would presumably include a lot of students who performed poorly on a given standardized test, meaning that the language ability range of the test takers in these studies is limited to the ones who performed well. Such a restricted range of ability might lead to limited power to detect a relationship between college success and standardized test scores (Cho and Bridgeman, 2012; Bridgeman et al., 2016).

Another possible reason for the limited predictive validity of standardized language tests, which has received little attention until fairly recently, is that individual differences in test preparation strategies could blur the relationship between test scores and academic success (Ginther and Yan, 2017). Standardized tests typically use multiple-choice items to measure listening and reading, and productive tasks, such as writing a short essay, summarizing a text, or disagreeing with a position, to assess writing and speaking. It may be that certain students can prepare for particular tasks in ways that help them attain higher scores on them than their actual language abilities would suggest, while they are unable to do so with other tasks. For instance, some students may be able to practice multiple-choice test taking strategies that could aid them in achieving higher scores on listening and reading sections of a test than their abilities would dictate. On the other hand, they may not be able to prepare for speaking and writing tasks in ways that would help them achieve scores beyond their language skills. This could lead to uneven score profiles (a high score on one or more of the four skills and a low score on one or more of the other skills), and students with these unbalanced score profiles could have some scores, which are not very representative of their true language abilities. It could be these students (ones with scores that do not represent their true language abilities) scores that limit the predictive validity of standardized tests.

A few studies have provided evidence that for certain students, particularly ones with unbalanced score profiles, standardized proficiency test scores are less predictive of academic success than for students with balanced score profiles. For instance, further analysis by Bridgeman et al. (2016) showed that when students were grouped by linguistic background and academic discipline, and students with discrepant score profiles were excluded from the analysis, the relationship between test scores and grades increased, in some cases quite dramatically. For Chinese Business students, for example, the correlation of TOEFL iBT reading scores with GPA was found to change from .01 to .36 when test takers with uneven profiles were removed from the analysis.

A similar finding was also reported by Ginther and Yan (2017), who explored the predictive validity of TOEFL iBT for three Chinese undergraduate student cohorts (N=1,990), 2011, 2012, and 2013, in an American university. In 2011 and 2012, the minimum entry language requirement was an overall score of 80 on TOEFL iBT; however, in 2013 an additional requirement was a minimum score of 18 on the iBT speaking and writing sections. Although reported correlations between TOEFL iBT scores (overall and section) and GPA were generally weak for the three cohorts, with correlations ranging from .07 to .32, the directions of the correlations were found to change by enrollment year. For the 2011 and 2012 cohorts (when students in the study may have had scores lower than 18 on speaking and/or writing), speaking and writing scores had a positive relationship with GPA; however, negative correlations were found for listening and reading scores. By contrast, for the 2013 cohort (when all students had speaking and writing scores of at least 18), no negative correlations were found between test scores and GPA. Given that strong reading and listening skills are expected to facilitate rather than inhibit academic success, these results may indicate the existence of distinctive Reading/Listening versus Speaking/Writing score profiles among the 2011 and 2012 cohorts. In addition, the absence of negative correlations for the 2013 cohort, for whom admission language requirements included both overall and section cut scores on TOEFL iBT, lends support to the argument that the use of section cut scores in entry-level language requirements may enhance predictive validity of standardized proficiency tests.

In another study, Harsch et al. (2017) investigated the utility of TOEFL iBT for predicting academic success in a UK setting. The researchers examined the relationship between 504 graduate students' test scores and their grades, taking into account linguistic background, discipline, and post-entry language support. While reported correlations of TOEFL iBT scores with course grades were generally weak, with correlations ranging from .10 to .20, it was found that the correlations varied systematically by linguistic background and academic discipline. For instance, for Chinese students, TOEFL iBT total and speaking scores correlated weakly, but significantly, with course grades (r values around .31), whereas for German students, none of the correlations were significant. Likewise, for disciplines with a quantitative focus, iBT speaking, listening and overall scores had weak, but significant, relations with grades (correlations around

.20), while for disciplines with a social sciences focus, no significant relationship was found between test scores and course grades.

What research indicates about the relationship between standardized academic language proficiency tests and academic success has several implications for the use of these tests for admission and placement purposes in tertiary education. Given that these test scores are not strong predictors of academic success for all students, particularly students with unbalanced score profiles or from certain L1 backgrounds, higher education institutions should consider either not using standardized academic language proficiency test scores or using them along with other criteria when making admissions decisions. For the first alternative, institutions would be expected to develop their own language proficiency assessments for admission purposes since academic language ability is an essential criterion for admission decisions. For the second alternative, institutions would be expected to set minimum cut-scores for each of the four skills for admissions decisions and use section cut scores based on each of the skills to support decisions about students' post-entry language support needs. For example, to be granted admissions, a student would need to exceed a certain score in each of the four skill areas along with a certain overall score on a given language test. Such a practice would help limit the concern that students using particular test preparation strategies to obtain high scores would be inappropriately granted admissions. That is, test takers who are prepared to do well on multiple-choice tests would also need to do well on productive tasks (e.g., writing an essay) to gain admissions. Considering that students with similar total scores, but unbalanced score profiles, may not have the same language needs, the use of section scores for placement purposes may also assist stakeholders in making better judgments about a student's academic language needs.

Using Standardized Language Proficiency Tests for Placement Purposes

With increasing numbers of international students studying in English-speaking countries, many universities try to find the most effective and practical way to determine newly admitted students' additional language support needs. A common practice among these institutions is the use of standardized academic language proficiency test scores for ESL placement decisions. Ling et al. (2014) reviewed the websites of 152 higher education institutions, surveyed 80 institutions, and interviewed representatives of 24 ESL programs. The researchers reported that standardized academic language proficiency tests were widely used to inform placement decisions, with around half of the 50 4-year universities utilizing TOEFL iBT and/or IELTS scores for identifying students who might benefit from additional language support. For stakeholders, such a practice is attractive for at least two reasons. First, standardized test scores are readily available from the admissions process; no additional testing is needed and this could save substantial resources in many educational contexts. Second, test takers generally believe in standardized proficiency test scores as accurate indicators of their language ability. This means that there is likely to be little disagreement from them about who needs the additional language classes.

Although the use of standardized language tests for placement purposes is commonplace in higher education institutions, it might be unrealistic to expect these proficiency tests to be very accurate at placing students into levels of a particular language support program. Intended to provide a general indication of a test taker's language ability, standardized academic language proficiency tests are designed to assess how much ability and/or capability a test taker has in a given language. By contrast, placement tests are developed to separate test takers into levels of a pre-existing language support program with the purpose of maximizing the effectiveness of language instruction in homogenous classes (Cheng and Fox, 2017). Because many programs design their own language support curriculum around program-specific learning objectives, it is expected that materials and teaching practices used in these courses may not well align with a test not specifically designed for the curriculum. In this regard, standardized academic language proficiency tests would not be expected to serve effectively as placement tools (although they might be useful for identifying students that almost certainly do not need any language support and ones that almost certainly do).

The fact that standardized academic language proficiency tests are practical and generally accepted as an accurate measure of language ability among test takers is also not sufficient to support the use of these tests for placement decisions. To determine the extent to which these tests are appropriate for placement purposes in higher education contexts, empirical investigations with sound methodological designs are needed. Unfortunately, such investigations are quite limited (Fox, 2004, 2009, Arigoni and Clark, 2015). In addition, while locally conducted studies may provide initial insight into the usefulness of standardized tests for placement into language support courses, caution is warranted when interpreting results from these studies. This is because much of the reported research relies on an insufficiently validated in-house placement test to evaluate the effectiveness of a standardized proficiency test. That is, in most cases students are first placed into levels of a pre-existing language support course based on their scores on a locally developed placement test, and then their standardized academic language proficiency test scores are compared across the levels determined by the locally developed test. In fact, it may be that the locally developed test, itself, may not be very effective at separating test takers into appropriate levels of an ESL course. Other research has used teacher and student judgements to determine the effectiveness of placement decisions made based on standardized test scores. Unfortunately, these judgements are often not very objective and may not be very good indicators of appropriate placement, either.

An example of a study that relied on locally developed placement test scores as a measuring stick for a standardized test is that of Kokhan (2012), who examined the effectiveness of TOEFL iBT for separating students into three levels of an ESL writing course. Based on the results of an in-house writing placement test, 2,363 students were placed into three course levels. The students' TOEFL iBT total and section scores were then compared for the students placed at the three levels and the scores did not align very well. For example, for a student with an overall iBT score of 80, the chances of being placed into Level 1, 2, and being exempted were 40%, 50%, and 10%, respectively. The researcher concluded that no particular set of TOEFL iBT total or section scores could serve reliably as placement cut scores since their use resulted in a substantial proportion of misplacement compared to the placements determined by the locally developed placement test.

Examples of studies which have utilized teacher and/or student judgements to evaluate the effectiveness of standardized academic language proficiency test scores for placing students into appropriate classes are common. Arigoni and Clark (2015) examined the usefulness of IELTS as a placement tool in an English-medium university in Cairo. Two-hundred ninety-eight students were placed into two levels of an academic writing course based on their IELTS overall and writing section scores. The findings obtained from student and instructor questionnaires and interviews indicated that instructors considered nearly 15% of the students as misplaced, and around 25% of the students believed that they needed to study more than their peers to be successful, suggesting a perceived misplacement among the students. A similar result was reported by Fox (2009), who also used teacher judgement about placement as a criterion against which the standardized academic language proficiency test scores were evaluated. She investigated the appropriateness of these tests (i.e., TOEFL iBT and IELTS) for placement decisions at a Canadian university. Students (N = 261) were separated into three levels of an English for Academic Purposes (EAP) course based on their standardized test scores, and a considerable number were identified as misplaced by classroom teachers.

Research on the use of standardized academic language proficiency tests for placement purposes seems to suggest that these tests may not function effectively for separating students into levels of a pre-existing language support program. Given the main aims of proficiency tests, this finding makes some intuitive sense; however, it remains unclear as to the effectiveness of standardized academic proficiency tests for placement purposes, given the methodological limitations discussed above. On the other hand, it is obvious that when test users plan to use a test for a purpose for which the test was not developed, they should justify the appropriateness of the test for the intended purpose. It is possible that a standardized academic language proficiency test that does not serve well as a placement tool in a certain program might be more appropriate for placement decisions in another program, depending on the objectives of the curriculum in the program. Therefore, test users must consider how the objectives of a particular language support program would align with those of standardized academic language proficiency tests before making decisions about the use of these tests for placement purposes. Furthermore, if locally developed placement tests are themselves in need of validation, it would not be very appropriate to consider them as a criterion for standardized tests. Obviously, more research is needed to evaluate the effectiveness of standardized academic proficiency tests for placement purposes in higher education contexts. The findings of such research may guide stakeholders in interpreting and using these test scores for making placement decisions.

Using Locally Developed Tests Rather than Standardized Tests for ESL Placement

Given the concerns about the use of standardized academic language proficiency tests for placement into ESL courses, a number of higher education institutions have developed their own placement tests. There are several commonly touted benefits of using locally developed placement tests. First, they can be aligned with local curriculum objectives, meaning that what is assessed can represent what students are expected to learn in a particular language support program (Green, 2012). Students can be placed into appropriate levels of an EAP program based on their level of mastery of the content of the curriculum, and those who are unable to pass portions of the test designed to assess the content covered at a certain level could be justifiably placed into that level. Second, the use of locally developed tests may provide teachers who are familiar with student profiles, curriculum objectives, and local needs with a chance to contribute to test design, which may enhance the appropriateness of a test in a local context. These teachers often have a strong understanding of the content of the courses and the types of student profiles that can most benefit from these courses. Having their input in the test design may lead to assessments that can better identify students that are likely to benefit from a given language support course. Third, for certain groups of test takers, locally developed placement test scores may be more representative or valid than standardized academic language proficiency test scores. As was discussed earlier in this chapter, a possible reason that students have uneven score profiles on standardized tests is that they have spent a great deal of time preparing for the tests. For instance, they may be able to perform above their true ability levels on certain task types, such as multiple-choice, in standardized tests. Because the stakes of locally developed placement tests are not as high as those of standardized tests (being required to take an additional course is not nearly as important as not being accepted to the university at all), students would not be expected to prepare excessively for these tests. As a result, locally designed placement tests may yield scores that are more representative of the students' true ability. However, it could be that placement tests with tasks or formats that are not well understood by a particular group of test takers might result in scores that do not reflect a test taker's true ability. This is because students may get lower scores than expected due to a lack of understanding of what they are supposed to do on the test to demonstrate their true language ability. This so called, "test effect" should also be avoided by making sure students understand what they are expected to do to complete the tasks. Thus, familiarity with test tasks can lead to scores that represent a test taker's true ability, but too much practice on a given task type that is susceptible to a study effect could lead to overestimates of a test taker's ability. It would therefore seem that locally developed placement tests aligned with local curriculum objectives and with tasks familiar to test takers would function better for placement purposes than standardized academic language proficiency tests. However, the extent to which this is the case is not as clear as it seems.

Just like much of the research on standardized assessments is conducted or funded by companies that develop these tests, research on locally designed tests is often undertaken by the test developers themselves, who have a strong motivation for finding evidence to support their intended use. This potential source of bias should be considered carefully when interpreting findings obtained from these studies. It should also be recognized that many local contexts are unlikely to have the resources necessary to develop and administer a well-designed placement test and to process and report test scores appropriately. In fact, many institutions lack the necessary expertise for this purpose even if they did have the financial resources available. As a result, locally developed placement tests often fail to measure all aspects of the abilities that they should. For instance, because assessing speaking requires a great amount of educational, financial, and logistical resources (Ockey, 2017), many local placement tests neglect to assess it directly (Ling et al., 2014).

Research on the effectiveness of locally designed tests for ESL placement has yielded inconclusive results. While some studies suggest that locally developed tests may not be useful for placement purposes, others show that they could serve reasonably well for separating test takers into appropriate levels. Lee and Greene (2007) investigated the extent to which scores on a locally developed test predict ESL students' academic performance and language-related difficulties in degree courses. A hundred students with TOEFL iBT scores below 102 were administered an in-house placement test designed to assess academic listening, reading, and writing skills. The findings indicated no correlations between test scores and GPA, suggesting that the in-house placement test was a poor predictor of academic success. The analysis of teacher and student evaluations and their relationship with test scores also showed that teachers' evaluations of students' course performance had no relationship with their test scores, while students' self-evaluations correlated moderately with them. At a university in the UK, Fulcher (1997) examined the validity and reliability of an in-house placement test designed to identify students in need of additional language support. The test, which included reading, descriptive and argumentative essay writing, and English structure sections, was administered to 1,619 students. Reported reliability indices were quite low for the reading and grammar sections, .63 and .59, respectively, and further analysis indicated that the grammar section was not sensitive enough to separate the students into appropriate levels of ESL courses. Writing, however, was reported to be quite reliably assessed, with an estimated reliability of .89.

Other studies have reported more positive results about the effectiveness of locally developed placement tests. For example, Winke (2013) investigated the usefulness of an in-house group oral test for placing students into seven levels of a 6-week ESL speaking course. Two group oral tasks, in which test takers were expected to have a discussion with their partners in a group, were administered to 128 students. Reported inter-rater reliability was .72, suggesting a moderate level of agreement between two raters. The findings of cluster analysis showed that the group oral test effectively placed students into the seven levels of the course.

To recap, although locally developed placement tests would at first glance appear to be more appropriate for placement decisions than standardized tests, this is not necessarily the case. The advantages of curriculum alignment, teacher input, and lack of a test preparation effect may be offset by a lack of local resources necessary for developing effective assessments. When local resources are sufficient to develop and sustain an effective test, however, it would seem that a local placement test may be more appropriate for making placement decisions than a standardized assessment that has not been designed for a particular local context.

Conclusions

This chapter began with a discussion about stakeholders' views and knowledge about interpretation and use of standardized academic language proficiency test scores. It was concluded that many stakeholders lack sufficient knowledge about these tests to use them appropriately in their own local contexts. To help stakeholders make informed decisions regarding the use of these tests in higher education contexts, it was argued that it is important to build assessment literacy among stakeholders. Efforts to increase assessment literacy among stakeholders, particularly policy makers, administrators, and instructors, are crucial since they are the ones most likely to make decisions about how the test scores are used. Additionally, test takers themselves should be better educated about the effectiveness of these tests for particular purposes, since their satisfaction or dissatisfaction with the use of standardized tests scores for particular purposes (as evidenced by complaints or silent “acceptance”) can affect the way score users decide to use these test scores.

Next, the usefulness of standardized language tests for predicting academic success was explored. It was shown that academic success is not very accurately predicted by standardized language test scores, and it was suggested that this was likely due to a plethora of reasons, including academic motivation, personality factors, homogeneous groups of test takers used in most studies, and possible negative effects of test preparation strategies employed by certain groups of test takers. These test preparation strategies may lead to not only unbalanced score profiles, but also scores not representative of test takers’ actual ability, particularly on certain task types. It was suggested that when standardized test scores are used for admission and/or placement decisions, minimum scores in each of the reported score categories be set to limit a possible study effect on decisions made based on test scores.

The use of standardized academic language proficiency test scores to place students into appropriate levels of language support courses and the alternative, the use of locally developed tests for this purpose, was then examined. It was concluded that standardized academic language proficiency tests may not be appropriate for placement into ability-based classes driven by a local curriculum. In such situations, it may be better to use a locally designed placement test aligned with the objectives of the curriculum as long as educational and financial resources are available. However, it was argued that in many higher education institutions, the resources and expertise necessary to develop and administer such tests effectively are unavailable. It was recommended that in such cases, stakeholders must first consider the degree of alignment between the objectives of a local curriculum and those of a standardized academic language proficiency test. If there is a high correspondence between the two, then the standardized test scores might be used with caution along with other proven ability indicators. It was also suggested that setting minimum cut-scores in each of the skill areas may also limit misplacements when standardized academic language proficiency test scores are used for placement decisions.

Standardized academic language proficiency tests have an important role in helping stakeholders make decisions about a test taker's academic language ability in tertiary education. When it is determined that the scores can be used for making a particular decision, it is important that they are used appropriately. They cannot be considered appropriate for all situations and infallible for any context. When their use can be justified empirically, they should be used in conjunction with other indicators of language ability and interpreted appropriately for the given context.

References

Arrigoni E, Clark V (2015) Investigating the appropriateness of IELTS cut-off scores for admissions and placement decisions at an English-medium university in Egypt. IELTS Research Reports Online Series 3: 1-29

Bachman L, Palmer A (2010) Language assessment in practice. Oxford University Press, Oxford

Baker, B (2016) Language assessment literacy as professional competence: The case of Canadian admissions decision makers. Canadian Journal of Applied Linguistics, 19(1), 63-83

Baker B, Tsushima R, Wang S (2014) Investigating language assessment literacy: Collaboration between assessment specialists and Canadian university admissions officers. Language Learning in Higher Education 4(1): 137-157

Bridgeman B, Cho Y, DiPietro S (2016) Predicting grades from an English language assessment: The importance of peeling the onion. Language Testing 33(3): 307–318

Carr N T (2011) Designing and analyzing language tests. Oxford University Press, Oxford

Cheng L, Fox J (2008) Towards a better understanding of academic acculturation: Second language students in Canadian universities. Canadian Modern Language Review 65(2): 307-333

Cheng L, Fox J (2017) Assessment in the language classroom: Teachers supporting student learning. Palgrave Macmillan, England

Cho Y, Bridgeman B (2012) Relationship of TOEFL iBT scores to academic performance: Some evidence from American universities. Language Testing 29:421–44

Coleman D S, Starfield S, Hagan, A (2003) The attitudes of IELTS stakeholders: Student and staff perceptions of IELTS in Australia, UK and Chinese tertiary institutions. IELTS Research Reports Series 4:161-235

Davies A, Brown, A, Elder C, Hill K, Lumley T, McNamara T (1999) Dictionary of Language Testing. Cambridge University Press, Cambridge

Dooey P, Oliver R (2002) An investigation into the predictive validity of the IELTS test as an indicator of future academic success. Prospects 17(1): 36–54

Fox J (2004) Test decisions over time: tracking validity. Language Testing 21(4): 437–465

Fox J (2009) Moderating top-down policy impact and supporting EAP curricular renewal: Exploring the potential of diagnostic assessment. Journal of English for Academic Purposes 8(1): 26–42

Fox J, Cheng L, Zumbo B D (2014) Do They Make a Difference? The Impact of English Language Programs on Second Language Students in Canadian Universities. TESOL Quarterly 48(1):57-85

Fulcher G (1997) An English language placement test: Issues in reliability and validity. Language Testing 14(2): 113-139

Ginther, A., & Elder, C. (2014). A comparative investigation into understandings and uses of the English Language Testing Service (Academic) Test, and the Pearson Test of English for Graduate Admissions in the United States and Australia: A case study of two university contexts. TOEFL iBT Research Report 24. Educational Testing Service, New Jersey

Ginther A, Yan X (2017) Interpreting the relationships between TOEFL iBT scores and GPA: Language proficiency, policy, and profiles. Language Testing. doi:10.1177/0265532217704010

Green, A (2012) Placement testing. In Coombe C, O' Sullivan B, Davidson P, Stoynoff S (eds) The Cambridge guide to language assessment. Cambridge University Press, Cambridge, pp 164–170

Harsch C, Ushioda E, Ladroue C (2017) Investigating the predictive validity of TOEFL iBT test scores and their use in informing policy in a United Kingdom university setting. TOEFL iBT Research Report 30. Educational Testing Service, New Jersey

Hyatt, D, Brooks, G (2009) Investigating stakeholders' perceptions of IELTS as an entry requirement for higher education in the UK. IELTS Research Reports, 10, 17-68

Inbar-Lourie O (2008) Constructing a language assessment knowledge base: A focus on language assessment courses. Language Testing 25(3): 385–402

Inbar-Lourie O (2013) Language assessment literacy. In Chapelle C A (ed) The encyclopedia of applied linguistics Blackwell, Oxford, pp 2923–2931

Kerstjens M, Nery C (2000) Predictive validity in the IELTS test: a study of the relationship between IELTS scores and students' subsequent academic performance. IELTS Research Report 3: 85–108

Kokhan K (2012) Investigating the possibility of using TOEFL scores for university ESL decision-making: Placement trends and effect of time lag. Language Testing 29(2): 291–308

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

Ling G, Wolf K M, Cho Y, Wang Y (2014) English-as-a-Second-Language Programs for Matriculated Students in the United States: An Exploratory Survey and Some Issues. ETS Research Report 14(11). Educational Testing Service, New Jersey

Messick S (1996) Validity and washback in language testing. Language Testing 13: 241–256

Ockey G J (2017) Approaches and challenges to assessing oral communication on Japanese entrance exams. JLTA Journal 20: 3-14

O'Loughlin K (2008) The use of IELTS for university selection in Australia: A case study. IELTS Research Reports 8: 2–98

O'Loughlin K (2011) The interpretation and use of proficiency test scores in university selection: How valid and ethical are they? Language Assessment Quarterly 8: 146–160

O'Loughlin K (2013) Developing the assessment literacy of university proficiency test users. Language Testing 30(3): 363–380

Pill J, Harding L (2013) Defining the language assessment literacy gap: Evidence from a parliamentary inquiry. Language Testing 30(3): 381–402

Popham W J (2004) Why assessment illiteracy is professional suicide. Educational Leadership 62(1): 82–83

Rea-Dickins P R, Kiely R, Yu G (2007) Student identity, learning and progression: The affective and academic impact of IELTS on 'successful' candidates. IELTS Research Reports 7. British Council, Australia

Riazi M (2013) Concurrent and predictive validity of Pearson Test of English Academic (PTE Academic). Papers in Language Testing and Assessment 2(2): 1-27

Taylor L (2009) Developing assessment literacy. Annual Review of Applied Linguistics 29: 21–36

Taylor L (2013) Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections. Language Testing 30(3): 403–412

Ushioda E, Harsch C (2011) Addressing the needs of international students with academic writing difficulties: Pilot Project 2010/11. Strand 2: Examining the predictive validity of IELTS scores (internal report). https://warwick.ac.uk/fac/soc/al/research/groups/llta/research/past_projects/strand_2_project_report_public.pdf. Accessed 31 Jan 2018

Winke P (2013) The effectiveness of interactive group orals for placement testing. In Second language interaction in diverse educational contexts (K. McDonough & A. Mackey, Eds) 247-268