SHIKEN

Volume 22 • Number 2 • December 2018

Contents

Testing and Evaluation SIG Newsletter

ISSN 1881-5537

Shiken

Volume 22 No. 2
December 2018

The TOEFL (ITP): A survey of teacher perceptions

John B. Collins and Nicholas H. Miller
john.buchanan.collins@gmail.com, millernicholash@gmail.com
Ritsumeikan Asia Pacific University

Abstract

We conducted a survey to develop an understanding of teachers' perceptions of the TOEFL (ITP) and its place within the English language program of a private university in western Japan. Our literature review reflects upon the history of the TOEFL and includes consideration of how three major factors underlying the recent paradigm shift in ELT assessment have contributed to its development. While the results of our survey display general agreement amongst teachers about the importance of the test for students on the program, they equally highlight misgivings about the TOEFL (ITP) as a measure of communicative competence that are consistent with this paradigm shift. Given the low level of agreement among teachers in the survey regarding the validity of the TOEFL (ITP), further investigation into teachers' perceptions of assessment could help to shed light on what factors underpin teachers views of the validity of the TOEFL (ITP) and language assessment in general. The results of our survey showed the lowest level of agreement among teachers in terms of the alignment of the TOEFL (ITP) with their beliefs about language teaching, as well as varied levels of willingness to teach the TOEFL (ITP), suggesting divergent views of the TOEFL (ITP) amongst teachers. We suggest that further research and discussion of this situation in relation to the broader construct of washback could help to inform decisions surrounding the TOEFL (ITP) and its place in the university's English program, as well as helping to advance our understanding of assessment washback in and beyond the context within which the present study took place.

Keywords: TOEFL (ITP), language assessment, EFL/ELT teacher perceptions, communicative competence, washback.

Originally conceived to address the issue of how to assess the English language proficiency of an increasing number of foreign students seeking to attend American universities, the TOEFL has become a widely held and highly influential assessment. To date, an estimated 35 million people worldwide have taken the TOEFL (ETS Global, 2012), and the high-stakes nature of the test has spawned a vast and lucrative TOEFL preparation industry. Universities often include some iteration of the TOEFL as part of their assessment framework, however the test has been subject to considerable scrutiny and criticism, particularly in relation to the paradigm shift that has taken place in language assessment in the field of ELT since the early 1990s. This situation raises the question of how teachers who are called upon to teach the TOEFL as part of their work themselves perceive the test. This article will explore teachers' perceptions of the TOEFL in relation to six issues, and examine the extent to which there is agreement among teachers in relation to these issues. Finally, we will reflect upon how the results of this study might be built upon through further investigation.

Assessment: Terminology and Definitions

The term assessment is generally agreed to embrace "a wider set of parameters than the term testing" (Rea-Dickins & Germaine, 1992 p. 3), but it is also worthwhile clarifying its relationship to the term evaluation, alongside which it is often discussed (e.g. Astin & Antonio, 2012). Assessment and evaluation are distinct but overlapping areas whose relationship is fundamentally defined by their respective focuses on individual learners and institutional issues. Lynch (2003), in an instructive formulation, employed the more specific terms *language assessment* and *program evaluation*. Language assessment, with its focus on the teaching and learning process as it affects individual learners, was defined as "the range of procedures used to investigate aspects of individual language learning and ability, including the measurement of proficiency, diagnosis of needs, determination of achievement in relation to syllabus objectives, and analysis of ability to perform specific tasks" (Lynch, 2003, p. 1). Program evaluation, on the other hand, concerned as it is primarily with institutional issues, and motivated to a large extent by

accountability requirements, was defined as "the systematic inquiry into instructional sequences for the purpose of making decisions or providing opportunity for reflection and action" (p. 1).

A Paradigm Shift in Language Assessment

Davison and Cummins (2007) characterized ELT as a field that for much of its history saw language assessment as the responsibility of specialists, divorced from the business of teaching and learning; "taken for granted…but often misunderstood by practitioners, rarely included as a component in English language teacher training, and never really challenged by key stake-holders" (p. 415). As a consequence, ELT lagged behind the rest of the educational field in exploring new theories and methods of assessment. However, since the early 1990s, a major paradigm shift in ELT assessment has taken place, driven by three major factors: a rise in expectations and forms of accountability as a result of economic restructuring and globalization; a major questioning of traditional forms of testing; and shifts in our constructs of language and language learning (Davison & Cummins, 2007).

As we shall see in the subsequent section, these factors have impacted upon the development of the TOEFL. It is also worth noting the broader effects of this paradigm shift on teachers around the world, for example in the tension that has emerged between standardized, high-stakes testing and teacher-based assessments, which has become a focal point around which controversy and conflict over competing visions of ELT as an enterprise tend to manifest (Davison & Cummins, 2007).

Language Assessment and the History of the TOEFL

Cumming (2007) drew attention to how ambitious and demanding it is to conceptualize, produce and administer a single test "to provide information that indicates whether people have achieved high levels of proficiency in a second language—and to do so comprehensively, validly, reliably, regularly, economically and fairly throughout the world" (p. 474). As such, it is not surprising that concerns, criticisms and reviews of the conceptualization of English language tests for university admissions have been many and varied. This is perhaps particularly true of the TOEFL, for some time now the most widely used of such tests, for whose development, operations and finances Educational Testing Services (ETS) has held sole responsibility since 1973, albeit with the interests of a range of groups concerned with the admission of international students to universities in North America being represented by the contemporaneously-formed TOEFL Policy Council (Taylor & Angelis, 2008).

The first major step in developing the TOEFL took place in Washington on May 11-12, 1961, at a conference sponsored by the Center for Applied Linguistics (CAL) in cooperation with the Institute of International Education (IIE) and the National Association of Foreign Student Advisers (NAFSA), that was called to address the issue of how to assess the English language proficiency of an increasing number of foreign students seeking to attend American universities (Spolsky, 1990). The participants grappled with what the test should measure, what content it should include, and how to justify its validity (Taylor & Angelis, 2008). The decisions regarding these issues, which gave shape to the TOEFL paper-based test (PBT), are illustrated in Table 1 (from Taylor & Angelis, 2008, p. 29).

However, a long-term project undertaken at ETS and subject to the advice and on-going review of the TOEFL Policy Council, initially called *TOEFL 2000*, which started in the early 1990s and whose ultimate goal was the development and validation of a new and improved TOEFL, constituted not only ETS's most systematic series of revisions, but the most comprehensive and methodical analyses of English proficiency tests for university admissions undertaken to date *per se* (Cumming, 2007). Informed by the paradigm shift mentioned in the preceding section, the *TOEFL 2000 Framework: A Working Paper* (Jamieson et al., 2000) outlined key areas of concern for those involved in the development of a new TOEFL. Firstly, the project acknowledged its accountability to its stakeholders: "the diverse constituencies served by the

TOEFL program" (p. 1). Foremost amongst these were college and university admissions officers, teachers of English as a second or foreign language, and examinees. These groups required, respectively: an efficient way to ascertain the proficiency levels of large numbers of international students; more detailed information to guide decisions regarding course placement and instructional design; and results that facilitate both reflection on students' progress in instructional settings and potential decisions on whether to apply to other programs (Jamieson et al., 2000).

Table 1.
Decisions about Testing for the First TOEFL

Testing Issue	*Decisions*
Construct: What should the test measure?	The test should consist of an "omnibus battery testing a wide range of English proficiency and yielding meaningful (reliable) subscores in addition to total scores" (*Testing the English Language Proficiency of Foreign Students*, 1961, p. 3). The construct can be defined as a list of components of language knowledge and skills that would be expected to affect performance across a wide range of relevant situations.
Content: What should the test consist of?	The test battery will contain sections for the measurement of: (a) control of English structure, (b) auditory comprehension, (c) vocabulary and reading comprehension, and (d) writing ability. All sections will consist of multiple-choice items.
Validation: How can test use be justified?	The test should be pretested on a wide range of nonnative speakers of English as well as native speakers of English. Correlations between the new test and existing tests should be obtained and reported.

(from Taylor & Angelis, 2008, p. 29)

Secondly, the project acknowledged the need to take into account the questioning of traditional forms of testing, particularly the TOEFL's association by many in the language teaching and testing communities with discrete-point testing, which, along with the exclusive use of traditional, multiple-choice items, was seen as having a negative impact on instruction (Jamieson et al., 2000). The TOEFL (PBT) was grounded in the discrete-point construct of language proficiency (Enright, 2018) and was characterised by the principles of what Spolsky (1975, as cited in Morrow, 1979, p. 144) described as the "psychometric-structuralist" era of language testing. This period of testing, also referred to by Morrow (1979) as the "Vale of Tears" *(ibid.),* was dominated by the ideas of the hugely influential Robert Lado, who championed an atomistic and behaviourist view of language acquisition and the use of multiple-choice questions to ensure objectivity and reliability. Criticism of this approach centred on the underlying assumptions about the nature of language acquisition and language knowledge. Lado (1961) grounded his ideas in the contrastive analysis approach of Charles C. Fries (1945) and the identification of *language problems*, namely, structural differences between the native language and the target language. Lado (1961) claimed that language acquisition is essentially a matter of mastering these problems to the point of habit and asserted that language tests should "attempt to test mastery of the units and patterns that are different from those of the native language and constitute the learning problems" (Lado, 1964, p. 165). This approach to language testing, as Morrow pointed out in his call for the profession to move beyond the "Vale of Tears" towards a "Promised Land" characterised by a communicative approach to language teaching and testing:

> depends utterly on the assumption that knowledge of the elements of a language is equivalent to knowledge of the language… Knowledge of the elements of a language in fact counts for nothing unless the user is able to combine them in new and appropriate ways to meet the linguistic demands of the situation (Morrow, 1979, p.144).

The third—and main—focus for development of a working framework for the revised TOEFL was to address calls by various constituencies for a test that was more reflective of current theories of communicative language use in an academic context (Jamieson et al., 2000). Calls for a test that accounts for an integrated and communicative model of language proficiency date back to Carroll's (1961) widely-cited recommendation that the TOEFL include an "integrated, facile performance on the part of the examinee" (p. 318). To facilitate communicative competence becoming the test's underlying construct, Chapelle, Grabe and Berns (1997) focused on how to define "communicative language proficiency for academic life" (p. 1). Their model identified four components of communicative competence—sociolinguistic competence; grammatical competence; strategic competence; and discourse competence — whose relationship they expressed through recourse to Bachman's (1990) framework of *communicative language ability*, which consists of "both knowledge, or competence, and the capacity for implementing, or executing that competence in appropriate, contextualized communicative language use" (p. 84). Nonetheless, the difficulty of connecting research on communicative competence to a test framework resulted in a consensus that the development strategy should be "split by moving the current TOEFL test with some important design enhancements to an interim computer-based test (CBT) in 1998, while continuing to pursue the original vision of TOEFL 2000 within the Research Division of ETS" (Jamieson et al., 2000, p. 6). In 2005, a new test delivered at official test centers via the internet, and thus referred to as the *internet-based test* (iBT), was introduced, with the CBT being discontinued the following year (Enright, 2018). The validity research for the TOEFL (iBT) was guided by an argument-based approach (Enright, 2018), resulting in a validity argument which was presented in Chapelle, Enright and Jamieson (2008) "in technical terms to an audience of specialists" (Chapelle, 2008, p. 349); fundamentally its basis was "the backing that has been found for the assumptions underlying the inferences of domain definition, evaluation, generalization, explanation and extrapolation" (*ibid.,* p. 346). Chapelle also offered a single schematic intended to communicate the basis for TOEFL score use to a broader audience (see Figure 1).

In his review of the TOEFL (iBT), Alderson congratulated ETS for, on the whole, achieving their goals of developing a more task-centred approach to test design, and score interpretation based on communicative competence, through:

> a clearer focus on the academic environment, based on research into the language of academic tasks, careful prototyping, trialling, revisions…The inclusion of a compulsory speaking section, the integration of skills in numerous tasks, the use of longer written and spoken texts which are obviously more authentic and academic in nature, and…a radical reduction of focus on grammar (Alderson, 2009, p. 627).

In addition to the PBT and iBT described above, ETS also offers a variant of the TOEFL, which is referred to as the Institutional Testing Program (ITP). According to the ETS literature, the ITP is available to colleges, universities and other such providers of English language programs as a "convenient, affordable and reliable assessment of English-language skills" (ETS, 2018a, para 1). ETS operates two versions of the ITP: Level 1 (for English language students at the intermediate to advanced levels), and Level 2 (for students at the high-beginning to intermediate levels). The Level 1 test, the version currently employed at Ritsumeikan Asia Pacific University (APU), takes 115 minutes to complete and has a total of 140 multiple choice questions (ETS, 2018b). ETS described a number of ways in which institutions can use ITP test scores including as a placement test, for admissions to short-term, non-degree programs, and as contributing documentation for scholarships. However, since the test draws questions from previous TOEFL (PBT) tests, it is "not fully secure and should not be used for (general) admission purposes" (Tannenbaum & Baron, 2011, p.1). Given that all students in the APU Standard English Track are required to take the TOEFL (ITP) at various stages in the program and demonstrate an improvement in scores, it can be said that the ITP is employed for *progress monitoring*. Furthermore, since students are required to produce a score of 500+ by the completion of the Standard Track, it can be argued that it also serves as

an *exiting examination* by which students can demonstrate proficiency in English listening and reading (Official Guide to the TOEFL ITP Test, 2014, p. 1). A score of 500+ is also required for English-basis students to progress to the advanced English courses (English Teachers' Handbook, 2018).

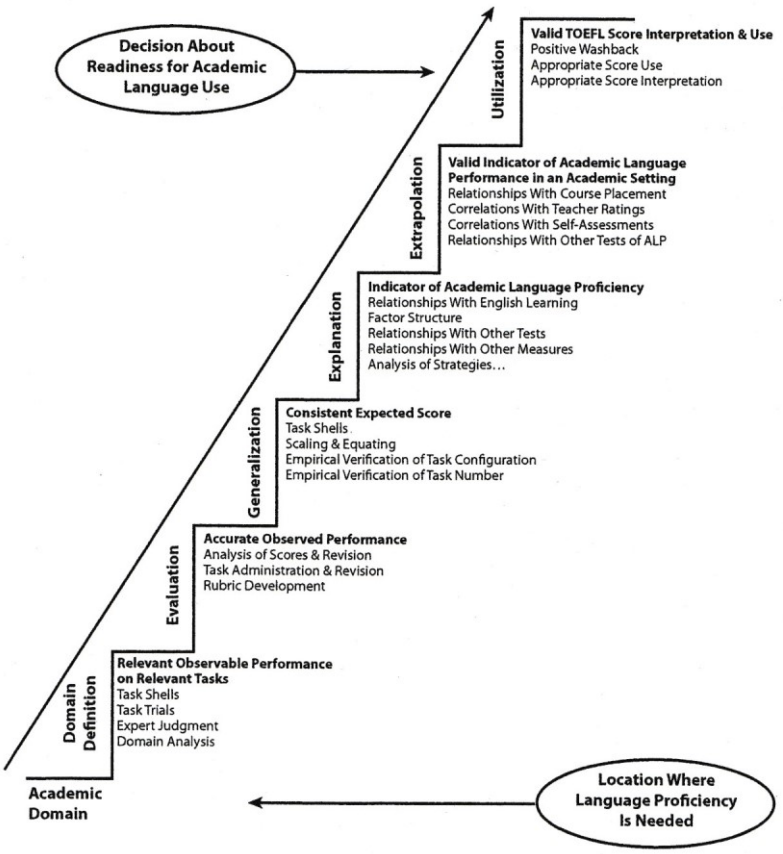

Figure 1: Steps of the TOEFL validity argument, with a diagonal line showing the argument moving toward valid score interpretation and use (from Chapelle, 2008, p. 349).

Method

Research Aims

By way of investigating the two research questions below, the aim of the current research was to establish whether, and to what extent, there is alignment among teachers' beliefs and views towards the primary high-stakes assessment used in the English Language Program, the TOEFL (ITP), which not only occupies a central position in the program's assessment framework, but has also been at the center of the paradigm shifts described above.

Research question 1. What are teachers' perceptions of the TOEFL (ITP) in relation to the following six issues: the importance of the TOEFL (ITP) for students; the alignment of the TOEFL (ITP) with the English program; the alignment of the TOEFL (ITP) with teachers' beliefs about English language teaching; willingness to teach TOEFL (ITP) classes; the TOEFL (ITP) as a measure of communicative competence; and the validity of the TOEFL (ITP) as a measure of English language proficiency?

Research question 2. To what extent is there agreement among teachers in relation to these six issues?

Research Context

This research was carried out at a mid-sized private university in rural western Japan. Established in 2000, the university now boasts a diverse student body of approximately 6,000, which is split 50-50 between international and domestic Japanese students. English language education is primarily carried out under the compulsory Standard Track, into which students are placed at levels ranging from elementary to upper-intermediate. Regardless of whether students study on an English or Japanese-language basis, they are required to complete a number of credits in the other language. Preparing students for this task is reflected in the program objectives:

> the…English program aims to cultivate in each student the English language knowledge and skills that they will need in order to communicate clearly and confidently with their fellow students, participate in lecture courses in English during their programs of study and use English in their working lives following graduation (English Teachers' Handbook, p. 2).

The vast majority of students in the English program are native Japanese speakers, in addition to a small number of international Japanese-basis students, predominantly Chinese and Korean. Students at the intermediate and upper-intermediate levels are required to take the TOEFL (ITP) at least once during the semester. The target TOEFL scores at the intermediate and upper-intermediate levels are 480 and 500 respectively. Prior to their completion of the Standard Track, students are therefore expected to clear a TOEFL (ITP) score of at least 500. In order to "maximize their scores and final grades, attention is given to test-taking strategies … in these courses and students are encouraged to work hard to attain the desired target score for their level(s)" (English Teacher's Handbook, p. 7). TOEFL skills classes are taught using the *Longman Preparation Course for the TOEFL* (Philips, 2003) and *The Complete Guide to the TOEFL Test: PBT Edition* (Rogers, 2011) and teachers are provided with a default lesson schedule that is arranged by test-taking skills such as "focusing on the second line" and "avoiding similar sounds". Students at the at the intermediate and upper-intermediate levels receive approximately twelve 95-minute TOEFL skills lessons which are embedded in the reading and vocabulary-focused B courses and constitute approximately half of the 15-week course. A considerable amount of class time is therefore spent by teachers preparing students for the TOEFL (ITP) although students are advised that approximately 200 study hours are required to move from a score of 435 to 470, and an additional 250 study hours to move from 470-504 (L. Stilp, personal communication, September 28, 2018).

Data-collection Instrument Development and Procedure

Considering the localized nature of the current research, a new self-report survey was designed that could shed light on the issues central to the TOEFL (ITP) and enable an informed discussion about the test and its place in the English program. The survey was also designed to reflect the recent paradigm shifts that have taken place regarding assessment validity and communicative competence (described above). Following the steps described in Dörnyei and Taguchi (2009), the survey was developed and focused on the following six aspects of the TOEFL (ITP) and its use within the English program:

1. The importance of the TOEFL (ITP) for students
2. Alignment of the TOEFL (ITP) with the English program
3. Alignment with beliefs about English language teaching
4. The TOEFL (ITP) as a measure of communicative competence
5. Willingness to teach TOEFL (ITP) classes
6. Validity of the TOEFL (ITP) as a measure of English language proficiency

Given that all the six aspects relate to subjective issues of attitude, belief and opinion, the survey was developed using multi-item scales, that is to say "cluster(s) of differently worded items that focus on the same target" (Dörnyei & Taguchi, 2009, p. 24). Items were worded as statements that respondents had to express their agreement or disagreement to via a 5-point Likert scale (1= *strongly disagree*; 5= *strongly agree*). Using the results of an initial trial, the internal reliability of each cluster was established by calculating Cronbach alpha coefficients. In order to produce a sufficiently high Cronbach alpha, each cluster was fine-tuned, including changing the wording and/or deletion of items (see Table 2 for final survey clusters and internal reliability coefficients; see appendix for full list of survey clusters and items).

Table 2
Breakdown of Survey Clusters and Respective Cronbach Alpha Coefficients

Cluster	Item no.	α
1. Importance of the TOEFL (ITP) for students	2, 6, 11, 19	.86
2. Alignment of the TOEFL (ITP) with the English program	3, 5, 16, 21	.80
3. Alignment with beliefs about English language teaching	4, 13, 17	.84
4. The TOEFL (ITP) as a measure of communicative competence	7, 10, 12	.84
5. Willingness to teach TOEFL (ITP) classes	8, 15, 20	.92
6. Validity of the TOEFL (ITP) as a measure of English language proficiency	1, 9, 14, 18	.82

Each cluster in the final version included at least three items – the recommended minimum number of items per cluster (Dörnyei & Taguchi, 2009). The final 21-item version was sent via Google Forms and completed by 19 English teachers who, at the time of the survey, either currently or had recently taught TOEFL (ITP) classes at the university.

Results

In order to answer research question 1, descriptive statistics were calculated for each cluster and are displayed in Table 3. For cluster 1, the mean was 4.45, indicating a high level of agreement that the TOEFL (ITP) was important for students. For clusters 2 and 3, the mean ranged between 3.03 and 3.00 respectively, indicating that teachers perceived a mid-to-low level of alignment of the TOEFL (ITP) with the English program, and with their beliefs about English language teaching. For cluster 4 (the TOEFL (ITP) as a measure of communicative competence) the mean was 2.39, indicating the lowest level of agreement of the six clusters. For clusters 5 and 6, the mean was 3.30 and 3.18 respectively, indicating a mid-to-low level of willingness to teach TOEFL (ITP) classes and mid-to-low level of agreement regarding the TOEFL (ITP) as a valid measure of English language proficiency.

In order to answer research question 2, descriptive statistics and box plots were generated using SPSS version 25 (see Figure 2). The level of agreement among teachers in relation to each of the six clusters was calculated as the interquartile range (the distance between the 25th and 75th percentiles) and is displayed in Figure 2 as the shaded area of the box plots. A higher level of agreement was indicated by a narrower interquartile range (IQR) and a lower level of agreement was indicated by a wider IQR.

Table 3
Means and standard deviations for the six clusters

Cluster	*M*	*SD*
1. The Importance of the TOEFL (ITP) for students	4.45	0.94
2. Alignment of the TOEFL (ITP) with the English program	3.03	0.89
3. Alignment with beliefs about English language teaching	3.00	1.04
4. The TOEFL (ITP) as a measure of communicative competence	2.39	0.94
5. Willingness to teach TOEFL (ITP) classes	3.30	1.25
6. Validity of the TOEFL (ITP) as a measure of English language proficiency	3.18	0.96

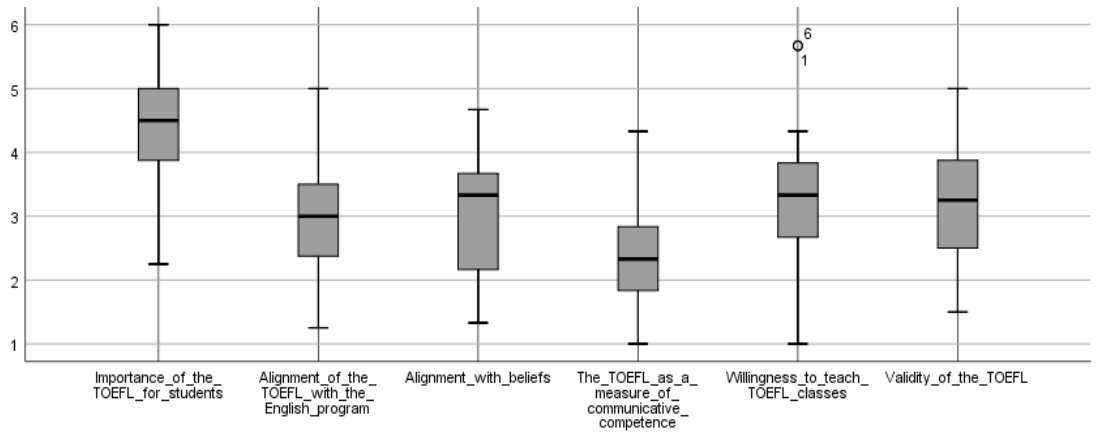

Figure 2: Box plots for the six clusters.

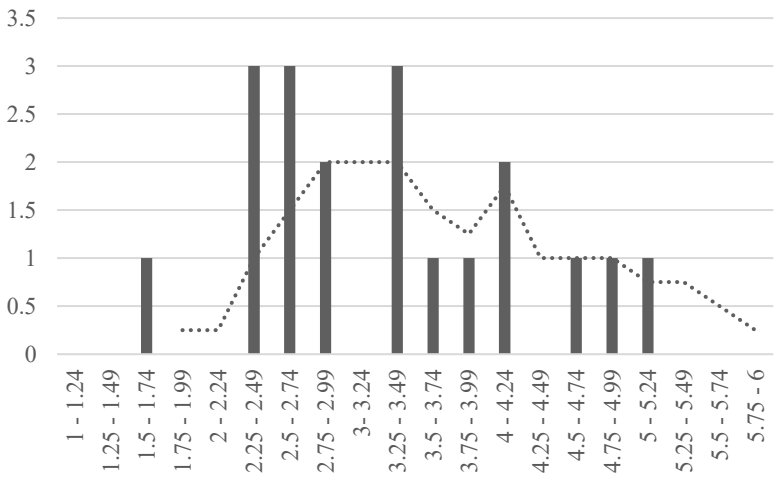

Figure 3. Results for cluster 6 (Validity of the TOEFL (ITP) as a measure of English language proficiency).

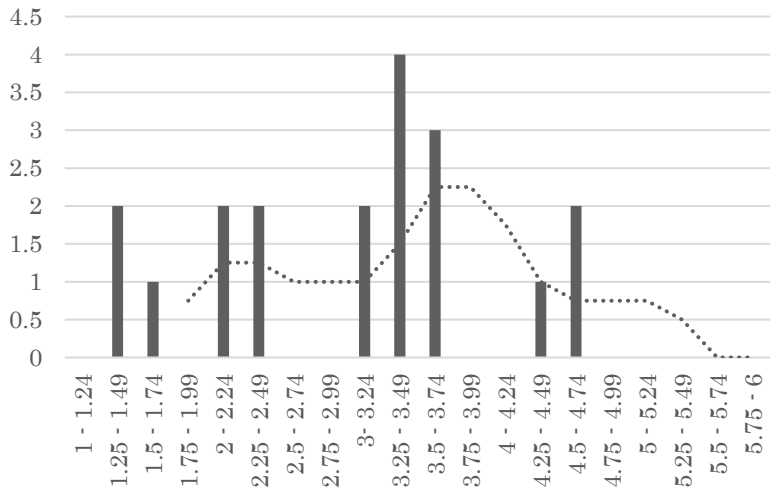

Figure 4. Results for cluster 3 (Alignment with beliefs about English language teaching).

Discussion

The results of our survey suggest that there is general agreement among teachers that the TOEFL (ITP) is important for students. This agreement could simply be a reflection of the high-stakes nature of the test within the English program and is not necessarily an endorsement of the test itself, which appears likely given teachers' responses to other survey items. Indeed, our survey results indicate that teachers have concerns about the TOEFL (ITP) that are consistent with the issues that drove the paradigm shift in ELT assessment that began in earnest in the 1990s, namely the major questioning of traditional forms of testing and shifts in our constructs of language and language learning (Davison & Cummins, 2007), which, in turn, were central to the *TOEFL 2000* project and the development of the TOEFL (iBT). The survey results point to teachers' misgivings about the TOEFL (ITP) as a measure of communicative competence, consistent with stakeholder calls underpinning the *TOEFL 2000* project for the development of a test that is more reflective of communicative competence models (Jamieson et al., 2000). Although it is beyond the scope of our current research, it is worth noting that while the ELT profession has moved toward a more integrated and performance-based approach to testing, communicative language testing gives rise to yet another set of concerns, particularly surrounding validity, authenticity and generalisability. Bachman (1990) described authenticity in relation to two approaches to language assessment, namely, the *real-life* (RL) approach and the *interactional/ability* (IA) approach—the former being primarily concerned with face validity and predictive utility, and the latter with construct validity. The RL approach, by seeking "to develop tests that mirror the 'reality' of non-test language use" (Bachman, 1990, p. 301), raises issues relating to what aspects of language knowledge to sample from the non-test domain and include in an assessment, and the generalizability of inferences; "Merely making an interaction 'authentic' does not guarantee that the sampling of language involved will be sufficient, or the basis for wide-ranging and powerful predictions of language behaviour in other situations" (Skehan, 1984, p. 208, as cited in Bachman, 1990, p. 311). There are also practical concerns to consider, as McNamara pointed out: "as assessment becomes more authentic, it also becomes more expensive, complex, and potentially unwieldly" (2000, p. 29). Considering the comparatively low level of agreement among teachers in our survey about the validity of the TOEFL (ITP), further investigation into teachers' perceptions of assessment validity could help to shed light on what factors underpin teachers' views of the validity of the TOEFL (ITP) and language assessment in general.

Our survey results showed a mid-to-low level of alignment of the TOEFL (ITP) with the English program and with teachers' beliefs about English language teaching, and a mid-to-low level of willingness among teachers to teach TOEFL (ITP) classes. A perceived lack of alignment of the TOEFL (ITP) with the English program suggests that use of the test is to some extent perceived as incompatible with achieving the stated aims of the course, which are: to cultivate in each student the English language knowledge and skills that they will need in order to communicate clearly and confidently with their fellow students; participate in lecture courses in English during their programs of study; and use English in their working lives following graduation (English Teachers' Handbook). As stated above, however, the English program also requires students to clear a TOEFL (ITP) score of at least 500 prior to their completion of the Compulsory Track. While it is evident that the TOEFL (ITP) aligns with the English program in so far as the program requires students to achieve a specific TOEFL (ITP) score, follow-up interviews with teachers could be helpful in identifying the specific reasons for this perceived lack of alignment with the broader aims of the English program.

The implications that this perceived lack of alignment raises might, in turn, be best discussed in relation to the broader construct of *washback*, defined by Messick as "the extent to which the introduction and use of a test influences language teachers and learners to do things they would not otherwise do that promote or inhibit language learning" (1996, p. 241). While the term washback has been described as "poorly defined" (Alderson & Wall, 1993, p. 117) and grounded in very little empirical evidence (Brown & Hudson, 1998), the notion that tests influence teaching is widely held (Alderson & Wall, 1993). Depending on whether it supports or hinders the achievement of educational goals, washback can be either positive or negative (Bailey, 1996; Alderson & Wall 1993) and can operate "in different ways in different situations" (Spratt, 2005). Alderson and Wall (1993) outlined 15 ways washback can theoretically influence teaching and learning, including *what* is taught and *how* it is taught, the *rate* and *sequence* of teaching and learning, and also the *degree* and *depth* of teaching and learning. Brown and Hudson (1998) described how a lack of alignment between an assessment and curriculum goals and/or objectives can result in negative washback, and offered the example of a multiple-choice test being employed at the end of a course that has communicative performance objectives.

As described above, since the goals of the English program emphasize developing the ability to *communicate* and *use* English, it could be argued that teachers' negative views toward the TOEFL (ITP) stem from a similar perceived mismatch between curriculum goals and assessment. On the other hand, given that achieving a TOEFL (ITP) score of 500 is a stated requirement for students and that TOEFL preparation classes constitute a substantial part of the English program during which teachers and students use a prescribed textbook focusing on test-taking skills, it is open to debate whether such classes and lesson content are examples of positive or negative washback; the exact nature of washback in this context is far from simple. Indeed, Alderson and Hamp-Lyons cautioned against simplistic interpretations of the washback hypothesis and suggested that tests "will have different amounts and types of washback on some teachers and learners than on other teachers and learners" (1996, p. 296). Based on their observation of TOEFL preparation classes and non-TOEFL preparation classes, they concluded that washback was largely mitigated by teachers' attitudes towards the test. Most teachers they spoke to described the TOEFL as boring and fragmentary, and resented the time pressures they felt they were under, while a small minority expressed positive views. While they did conclude that TOEFL preparation classes differed from non-TOEFL classes in terms of what and how teachers taught, Alderson and Hamp-Lyons (1996) suggested that the test alone was not the cause of washback, but rather that it is largely caused by the teachers' attitudes toward the test and the teaching decisions they make. The results of our survey showed the lowest level of agreement (of the six clusters) among teachers in terms of alignment of the TOEFL (ITP) and their beliefs about language teaching, suggesting divergent views of the TOEFL (ITP) amongst teachers. Teachers' willingness to teach TOEFL (ITP) classes (cluster 5) likewise elicited responses

ranging across the entire 1-6 Likert scale. Given this range of responses, the degree to which our survey respondents felt that the TOEFL (ITP) creates negative washback on their classes and the program, and how this manifests in their teaching decisions, would be a fruitful avenue of research to pursue, and could help to shed further light on the complex and interrelated factors that influence washback.

Conclusions

The aim of this research was to gain a better understanding of teachers' perceptions of the TOEFL (ITP) within the university's English program and to identify where there is agreement and divergence of opinion. The results of our survey indicate that teachers generally agree that the test is important for their students, and share doubts about the test as a measure of communicative competence that are consistent with previous concerns raised about the TOEFL (ITP). The greatest divergence among teachers' beliefs was found in regard to the alignment of the TOEFL (ITP) with their beliefs about English language teaching. As described above, this divergence of opinion, its connection to willingness to teach TOEFL classes and the potential for negative washback offers promising avenues for further research. On the one hand, the results of such research could help to inform decisions surrounding the TOEFL (ITP) and its place in the English program, including the assigning of teachers to teach TOEFL preparation classes based on the degree to which the test aligns with their beliefs about language learning and their willingness to teach such classes. In this way, the impact of negative washback could be reduced or avoided, learning outcomes for students improved, teacher job satisfaction raised, and curriculum goals potentially achieved. On the other hand, such results could also help to advance our understanding of assessment washback in and beyond the context within which the present study took place.

References

Alderson, J. C. (2009). Test review: Test of English as a foreign language: Internet-based test (TOEFL iBT). *Language Testing, 29*(4), 621-631. doi: 10.1177/0265532209346371

Alderson, J. C., & Hamp-Lyons, L. (1996). TOEFL preparation courses: A study of washback. *Language Testing*, *13*(3), 280-297. doi: 10.1177/026553229601300304

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics*, *14*(2), 115-129. doi: 10.1093/applin/14.2.115

Astin, A. W., & Antonio, A. L. (2012). A*ssessment for excellence: The philosophy and practice of assessment and evaluation in higher education.* Maryland: Rowman and Littlefield.

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

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

Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly*, *32*(4), 653-675. doi: 10.2307/3587999

Carroll, J. B. (1961). Fundamental considerations in testing for English language proficiency of foreign students. *Testing the English proficiency of foreign students.* (pp. 30-40) Washington DC: Center for Applied Linguistics.

Chapelle, C. A. (2008). The TOEFL validity argument. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson, (Eds.), *Building a validity argument for the test of English as a foreign language* (pp. 319-352). New York: Routledge.

Chapelle, C. A., Grabe, W., & Berns, M. (1997). *Communicative language proficiency: Definitions and implications for TOEFL 2000.* Princeton, NJ: ETS.

Cumming, A. (2007). New directions in testing English language proficiency for university entrance. In J. Cummins & C. Davison (Eds.), *International handbook of English language teaching* (pp. 473-485). New York: Springer.

Davison, C., & Cummins, J. (2007). Assessment and evaluation in ELT: Shifting paradigms and practices. In J. Cummins & C. Davison, C. (Eds.), *International handbook of English language teaching* (pp. 415-420). New York: Springer.

Dörnyei, Z., & Taguchi, T. (2009). *Questionnaires in second language research: construction, administration, and processing*. (2nd. Ed.), New York: Routledge.

English teachers' handbook (2018). Beppu, Japan: Ritsumeikan Asia Pacific University: Center for Language Education.

Enright, M. (2018). *TOEFL research insight series, volume 6: TOEFL program history*. Retrieved from https://www.ets.org/s/toefl/pdf/toefl_ibt_insight_s1v6.pdf

ETS (2018a). *About the TOEFL ITP assessment series*. Retrieved from https://www.ets.org/toefl_itp/about/

ETS (2018b). *Test content - The TOEFL ITP tests at a glance*. Retrieved from https://www.ets.org/toefl_itp/content/

ETS Global (2012). *The TOEFL iBT test*. Retrieved from https://www.etsglobal.org/Tests-Preparation/The-TOEFL-Family-of-Assessments/TOEFL-iBT-Test

Fries, C. (1945). *Teaching and learning English as a second language*. MI: University of Michigan Press

Jamieson, J, Jones, S., Kirsch. I., Mosenthal, P. & Taylor, C. (2000). *TOEFL 2000 framework: A working paper.* Princeton, NJ: ETS.

Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. London: Longman.

Lado, R. (1964). *Language teaching, a scientific approach*. London: McGraw-Hill

Lynch, B. K. (2003). *Language assessment and program evaluation.* Edinburgh: Edinburgh University Press.

McNamara, T. (2000). *Language testing (Oxford introduction to language series).* Oxford: Oxford University Press.

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*(3), 241-256. doi: 10.1177/026553229601300302

Morrow, K. (1979). Communicative language testing: Revolution of evolution? In C. K. Brumfit & K. Johnson, (Eds.), *The communicative approach to language teaching*. (pp. 143-159) Oxford: Oxford University Press.

Official guide to the TOEFL ITP test (2014). Educational Testing Service (ETS).

Phillips, D. (2003). *Longman preparation course for the TOEFL*. Longman.

Rogers, B. (2011). *The complete guide to the TOEFL test*. Heinle Cengage Learning.

Rea-Dickins, P., & Germaine, K. (1992). *Evaluation.* Oxford: Oxford University Press.

Skehan, P. (1984). Issues in the testing of English for specific purposes. *Language Testing*, *1*(2), 202-220. doi: 10.1177/026553228400100205

Spolsky, B. (1975). Language testing: Art of science. Main lecture delivered at AILA World Congress.

Spolsky, B. (1990). The prehistory of TOEFL. *Language Testing, 7*(1), 98-118. doi: 10.1177/026553229000700107

Spratt, M. (2005). Washback and the classroom: The implications for teaching and learning of studies of washback from exams. *Language Teaching Research*, *9*(1), 5-29. doi: 10.1191/1362168805lr152oa

Tannenbaum, R. J., & Baron, P. A. (2011). Mapping TOEFL ITP scores onto the Common European Framework of Reference. *Research Memorandum ETS RM*, pp. 11-33.

Taylor, C. A. & Angelis, P. (2008). The evolution of the TOEFL. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson. (Eds.), *Building a validity argument for the test of English as a foreign language* (pp. 27-54) New York: Routledge.

Appendix

Survey items (with clusters and item numbers)

Cluster 1: The importance of the TOEFL (ITP) for students
2. Students benefit from attaining a high TOEFL score.
6. It is beneficial for students that they achieve a high TOEFL score.
11. Achieving a high TOEFL score will help students in the future.
19. Reaching a high TOEFL score is important for our students.

Cluster 2: Alignment of the TOEFL (ITP) with the English program
3. The TOEFL fits well with the aims of the English program.
5. Teaching TOEFL aligns well with the goals of the English program.
16. The TOEFL fits into the English program well.
21. The TOEFL is inconsistent with the objectives of the English program

Cluster 3: Alignment with beliefs about English language teaching
4. Teaching TOEFL classes conflicts with my beliefs about language teaching.
13. TOEFL classes fit well with my beliefs about language teaching.
17. TOEFL classes align well with my beliefs about how languages are best taught.

Cluster 4: The TOEFL (ITP) as a measure of communicative competence
7. Communicative competence is reflected in students' TOEFL performance.
10. Communicative competence can be accurately measured by the TOEFL.
12. A high TOEFL score is a good indication of a high level of communicative competence.

Cluster 5: Willingness to teach TOEFL (ITP) classes
8. I am happy about teaching TOEFL classes now and in the future.
15. I prefer teaching classes that focus on the TOEFL.
20. I enjoy teaching TOEFL classes.

Cluster 6: Validity of the TOEFL (ITP) as a measure of English language proficiency
1. TOEFL scores provide an accurate indication of students' proficiency levels.
9. TOEFL scores closely align with the test takers' actual language proficiency.
14. TOEFL scores provide a valid measurement of English proficiency.
18. TOEFL scores provides students with an accurate picture of their English proficiency.

Questions and answers about language testing statistics:

Calculating reliability of dictation tests: Does K-R21 work?

James Dean Brown brownj@hawaii.edu
University of Hawai'i at Mānoa

Question:

For many tests like multiple-choice, true-false, and fill-in, we have item statistics which we can use in calculating reliability statistics like K-R20 and alpha. But for dictations, we only count-up total scores. So, my question is this: (a) can we use K-R21 based on the mean, standard deviation, and number of items for the total scores to calculate the reliability of a dictation, and (b) if so, how long should a dictation be in order to be reliable?

Answer:

This is the first of two columns that I will use to answer your questions. In the next column, I will discuss the relationship between dictation length and reliability. In this one, I will explore some problems and solutions for calculating the reliability of dictations. To do so, I will address four central questions:

1. What data serve as the basis for the current column?
2. What are some options for calculating reliability for dictations and what are the relationships among them?
3. What else is important in interpreting these common reliability estimates?
4. What does all this mean for calculating the reliability of dictation scores?

What data serve as the basis for the current column?

The participants were 220 graduate and undergraduate students taking the English Language Institute Placement Test (ELIPT) at UHM in fall 2015 and spring 2016. Their scores on the Internet-based TOEFL ranged roughly from 61 to 100.

Three dictations were involved here [For fuller descriptions, see Brown & Trace (2018).]:

1. The traditional academic dictation test (DCT) was a 50-word passage on a general academic topic that was recorded by a male native speaker of English who read three times: once at regular speed, once with pauses, and once again at regular speed. The 50 words were scored one point each. Spelling was not counted if the word was morphologically correct.
2. The connected-speech narrative (CSN) dictation was based on a passage about international student life in the US (from Prator & Robinett, 1972). It was recorded in much the same way as the DCT. However, the speaker was told to speak use connected speech just as in natural speech. While the passage contained 187 words, only the first 50 words involved in connected speech were scored as correct or incorrect. Connected speech was defined as changes from the dictionary pronunciation of the words including instances of adding, dropping, transitioning, or changing sounds. Otherwise, this dictation was scored the same as the DCT.

3. The connected-speech conversation (CSC) dictation was a 10-turn informal dialogue between a male and female about travel plans (based on Brown & Kondo-Brown, 2006), both speakers purposely used connected speech as appropriate. The CSC contained 83 words, 50 of which were scored because they involved connected speech. Otherwise, this dictation was scored the same as the DCT.

Note that we took the unusual step for all three dictations of compiling item level data, where each word was scored right or wrong and represented one item.

What are some options for calculating reliability for dictations and what are the relationships among them?

Since responsible interpretation of reliability estimates depends on descriptive statistics, Table 1 presents the mean, standard deviation (*SD*), minimum score, maximum score, and range for each of the three sets of dictation scores being considered here: the DCT, CSN, and CSC.

Table 1
Descriptive Statistics for the DCT, CSN, and CSC Dictations

Statistic	DCT	CSN	CSC
Mean	31.40	30.35	38.09
SD	8.34	8.84	6.58
Minimum score	11	14	16
Maximum score	50	50	50
Range	40	37	35

Notice in Table 1 that the CSC dictation has the highest mean at 38.09, and that the DCT and CSN were considerably lower at 31.40 and 30.35, respectively. This probably means that the conversational English in the CSC was easier for L2 learners to understand. Notice also that the *SD* for the CSC is considerably lower at 6.58, than those for the DCT (8.34) and CSN (8.84). This means that the scores on the CSC were less widely dispersed than those on the DCT and CSN. The maximum values of 50 indicate that at least one of the examines scored the highest possible score of 50 on each of the three dictations. The minimum values indicate the lowest scores were 11, 14, and 16, respectively. The range is another indicator of the relative spread of scores with the DCT being the widest (40) and the CSC the narrowest (35) and the CSN in between (37).

Table 2 shows four reliability estimates each for the DCT, CSN, and CSC dictations: (a) Kuder-Richardson formula 20 (K-R20); (b) Cronbach alpha; (c) Kuder-Richardson formula 21 (K-R21); and (d) split-half adjusted based on odd and even scores (for more on these various reliability estimates, see Brown, 2005, pp. 169-198, or Brown, 2016, pp. 107-153). Notice that K-R20 (.89, .90, & .87) and alpha (.89, .89, & .87) estimates are very similar for each of the dictations, which is what theory would predict. Notice also that the K-R21 estimates are systematically the lowest at .85, .87, and .81, respectively, and that the split-half adjusted estimates are the highest at .93, .92, and .91, respectively. In the next section, I will consider each of these four internal consistency reliability statistics in more detail.

Table 2
Four Types of Reliability Estimates for the DCT, CSN, and CSC Dictations

Reliability Estimate	DCT	CSN	CSC
K-R20	.89	.90	.87
Cronbach alpha	.89	.89	.87
K-R21	.85	.87	.81
Split-half adjusted (odd-even)	.93	.92	.91

What else is important in interpreting these common reliability estimates?

The *K-R20* formula (originally from Kuder & Richardson, 1937) is based on item- and test-level statistics, and it assumes "that the matrix of inter-item correlations has a rank of one and that all intercorrelations are equal" (p. 156). These assumptions are satisfied on a test where all items are measuring the same factor. Thus, unidimensionality is assumed. An additional design condition for K-R20 is that the items must be scored dichotomously (e.g., right or wrong).

Cronbach alpha is similar to K-R20 in that both are based on item- and test-level statistics, and alpha also assumes unidimensionality, but alpha has the advantage over K-R20 of being applicable to scales that are not dichotomous like weighted items (e.g., 0 = wrong, 1= partial credit, and 2 = completely correct), Likert items (e.g., 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, and 5 = strongly agree), etc.

The *K-R21* formula is based here on just three test-level statistics (the mean, standard deviation, and number of items), and it assumes unidimensionality, but also "that all items have the same difficulty" (Kuder & Richardson, 1937, p. 158). That last assumption is met if the item facility values on a test are approximately equal as on a test developed by selecting items from a pilot version that have item facility values ranging from .30 to .70 in item facility (i.e., the item difficulties are approximately equal) for a final version of the test (see Brown, 2005, pp. 66-68; Brown, 2016, pp. 63-67).

The *split-half adjusted* approach is based on scoring the odd-numbered and even-numbered items separately—producing two separate scores for each examinee and then calculating the correlation between the two sets of scores to find the half-test reliability and adjusting for full-test reliability using the Spearman-Brown prophecy formula (see Brown, 2005, pp. 177-179; Brown, 2016, pp. 125-130).

One way or another, all the internal consistency statistics discussed here estimate the degree to which items on a test are interrelated. And, one assumption statistically and logically of these estimates is unidimensionality, which means that the items on the test should all be measuring the same construct. While items on a test will most often be interrelated, that does not necessarily mean they are unidimensional as a set (as we will see below). So, in a sense, interrelatedness is a precondition, but is not sufficient in itself to assure unidimensionality. It would therefore be a mistake to be satisfied with a high degree of item interrelatedness (i.e., reliability) without also examining unidimensionality.

One way to examine the unidimensionality of a set of items is to run a factor analysis to see how many factors underlie what they are testing. A factor analysis—actually, a principle components analysis (PCA)—was performed for each of the three dictations with the 50 items in each case as the variables (for much more on factor analysis, see Brown, 2016, pp. 237-276). As shown in Table 3, the PCAs for DCT, CSN, and CSC produced 17, 16, and 18 Eigen values above 1.00, respectively, which accounted for 67.8, 67.8, and 72.4 percent of the variance, respectively. This is a clear indication that a number of components (or dimensions) underlie whatever these dictations are testing. Hence, all three seem to violate the assumption of unidimensionality, and as a result the reliability statistics found here probably provide

underestimates of the reliability of the scores. Such estimates are often referred to as lower-bound estimates of reliability, that is, the reliability of the scores will be at least as high as the estimate but may be higher.

Table 3
Item Facility, Eigen Values, and Percentage of Variance for the DCT, CSN, and CSC Dictations

Statistic	DCT	CSN	CSC
Number of Eigen values over 1.00	17	16	18
Percentage of variance accounted for	67.80	67.80	72.40
IF minimum	.10	.09	.05
IF maximum	1.00	1.00	1.00

Also recall that the K-R21 assumes that the item facility values on a test are approximately equal—say between .30 and .70. Table 3 indicates that the IF values ranged much more widely than that: from .10 to 1.00, .09 to 1.00, and .05 to 1.00, respectively. These violations of the assumption of equal difficulty may explain why the K-R21 estimates shown in Table 2 are consistently lower than all the other estimates. Brown (1983) reported similar but much bigger underestimates for K-R21 for cloze tests, where the assumption of equal difficulty is also violated.

Local item independence is seldom written about in language testing, but important for thinking about the reliability estimates for dictations. Essentially, many item and test statistics require that the items be independent, which is to say that they should not be correlated for reasons other than the fact that they are testing the same construct (for more on this, see Yen, 1993). This may be a problem when five items are based on the same reading or listening passage because being based on the same passage may cause items to be related for reasons other than the fact that they are testing the same construct. Local independence may also be a problem in cloze tests because answering one item correctly may help (for reasons beyond the construct being tested) answer the next blank because more context is available. The second study in Brown (1983, pp. 243-250) was an experiment that showed that, if lack of local independence is a problem for cloze procedures, it does not affect reliability.

Dictations may have the same problem of lack of local independence because writing down one word correctly may help (for reasons beyond the construct being tested) guess/know the next or other words because more context is provided. The net effect of dictations lacking local item independence might be that such reliability estimates would provide inflated estimates of the true state of affairs. This is troubling, and unfortunately, there is no research on this issue for dictations that I am aware of.

Interestingly, since the odd- and even-item scores used to calculate split-half adjusted reliability are more independent (i.e., items are not right next to each other) than the individual items (right next to each other) used in calculating the K-R20, alpha, and K-R21, I would expect any such inflation of the reliability estimates to affect the split-half adjusted estimates to a lesser degree than the others. Yet, it turned out that the split-half adjusted estimates were higher than the others. Thus, it appears that the magnitude of any inflation due to lack of independence, may be less than the magnitude of any underestimation due to lack of unidimensionality. This conclusion is based on relatively small data sets and a small number of dictations, so further research is clearly warranted before accepting any such conclusion.

For testers who are worried about this issue, a test-retest or parallel-forms reliability estimates (see Brown, 2005, pp. 175-176) would get around this problem—at least for statistically estimating reliability.

What does all this mean for calculating the reliability of dictation scores?

So, what are language testers to do if they want to know how reliable the scores on their dictations are? If they are concerned about lack of unidimensionality, the internal consistency estimates used here will work fine, but must be interpreted as underestimates. In any case, interpretation should be done cautiously (as in Brown, Phung, Hsu, Trace, Harsch, & Faucette, 2018, p. 24, where we put a footnote in the table containing our DCT reliability estimate as follows: "**K-R21 = very rough estimate"). In addition, the fact that dictations are clearly measuring multiple dimensions must be considered.

If language testers are concerned about lack of local independence, they could use either a test-retest or parallel-forms approach to calculate reliability because those approaches are based on independent total scores. These approaches involve considerably more work (for both the tester and examinees) than any of the internal consistency reliability estimates reported here, but they do avoid the local independence problem.

Conclusion

In this column, I (a) described the data that serve as the basis for the discussion in this column; (b) provided some options for calculating the reliability of dictations and looked at the relationships among them; (c) considered other important issues involved in interpreting these reliability estimates; and (d) suggested what all this means for calculating the reliability of dictation scores. In the process, for the internal consistency estimates presented here, it became clear that (a) they likely violate the assumption of unidimensionality and (b) that they also lack local independence across items. I ended by suggesting strategies for dealing with both problems. In direct answer to our question, yes, you can use K-R21, but only very cautiously while taking into account the issues discussed in this column.

I hope this column addressed the first part of your question adequately and provided you with the information you need for calculating and describing the reliability of your dictation tests in the future. [For much more on calculating, describing, and using reliability estimates see Brown (2005, pp. 169-198).] In the next column, I will explore the relationship between reliability and dictation length.

References

Brown, J. D. (1983). A closer look at cloze: Validity and reliability. In J.W. Oller, Jr. (Ed.) *Issues in Language Testing Research* (pp. 237-250). Rowley, MA: Newbury House (also available from ERIC: ED 227 695).

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment* (New edition). New York: McGraw-Hill.

Brown, J. D. (2016). *Statistics corner: Questions and answers about language testing statistics.* Tokyo: JALT.

Brown, J. D., & Kondo-Brown, K. (2006). Testing reduced forms. In J. D. Brown & K. Kondo-Brown (Eds.), *Perspectives on teaching connected speech to second language speakers* (pp. 247-264). Honolulu, HI: University of Hawaii, National Foreign Language Resource Center.

Brown, J. D., & Trace, J. (2018). Connected-speech dictations for testing listening. In G. Ockey & E. Wagner (Eds.). *Assessing L2 listening: Moving towards authenticity* (pp. 45-63). Philadelphia, PA: John Benjamins

Brown, J. D., Phung, H., Hsu, W-L, Trace, J., Harsch, K., & Faucette, M. P. (2018). 2016-2017 English Language Placement Test (ELIPT) revision project. *Second Language Studies, 36*(2), 1-25.

Kuder, G. F. & Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometrika, 2*, 151-160.

Prator, C. H., & Robinett, B. W. (1972). *Manual of American English pronunciation*. New York: Holt, Rinehart, & Winston.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187-213.

Where to submit questions:

Your question can remain anonymous if you so desire. Please submit questions for this column to the following e-mail or snail-mail addresses:

brownj@hawaii.edu.
JD Brown
Department of Second Language Studies, University of Hawai‘i at Mānoa
1890 East-West Road
Honolulu, HI 96822 USA

Call for Papers

Shiken is seeking submissions for publication in the June 2019 issue. Submissions received by 1 March, 2019 will be considered, although earlier submission is strongly encouraged to allow time for review and revision. *Shiken* aims to publish articles concerning language assessment issues relevant to classroom practitioners and language program administrators. This includes, but is not limited to, research papers, replication studies, review articles, informed opinion pieces, technical advice articles, and qualitative descriptions of classroom testing issues. Article length should reflect the purpose of the article. Short, focused articles that are accessible to non-specialists are preferred and we reserve the right to edit submissions for relevance and length. Research papers should range from 4000 to 8000 words, but longer articles are acceptable provided they are clearly focused and relevant. Novice researchers are encouraged to submit, but should aim for short papers that address a single research question. Longer articles will generally only be accepted from established researchers with publication experience. Opinion pieces should be of 3000 words or less and focus on a single main issue. Many aspects of language testing draw justified criticism and we welcome articles critical of existing practices, but authors must provide evidence to support any empirical claims made. Isolated anecdotes or claims based on "commonsense" are not a sufficient evidential basis for publication.

Submissions should be formatted as a Microsoft Word (.doc or .docx format) using 12 point Times New Roman font, although plain text files (.txt format) without formatting are also acceptable. The page size should be set to A4, with a 2.5 cm margin. Separate sections for tables and figures should be appended to the end of the document following any appendices, using the section headings “Tables” and “Figures”. Tables and figures should be numbered and titled following the guidelines of the *Publication Manual of the American Psychological Association, Sixth Edition*. Within the body of the text, indicate approximately where each table or figure should appear by typing “Insert Table x” or “Insert Figure x” centered on a new line, with “x” replaced by the number of the table or figure.

The body text should be left justified, with single spacing for the text within a paragraph. Each paragraph should be separated by a double line space, either by specifying a double line space from the Microsoft Office paragraph formatting menu, or by manually typing two carriage returns in a plain text file. Do not manually type a carriage return at the end of each line of text within a paragraph.

Each section of the paper should have a section heading, following the guidelines of the *Publication Manual of the American Psychological Association, Sixth Edition*. Each section heading should be preceded by a double line space as for a regular paragraph, but followed by a single line space.

The reference section should begin on a new page immediately after the end of the body text (i.e. before any appendices, tables, and figures), with the heading “References”. Referencing should strictly follow the guidelines of the *Publication Manual of the American Psychological Association, Sixth Edition.*