JALT Journal

JALT Journal is the research journal of the Japan Association for Language Teaching (JALT). It is published semiannually, in May and November. As a nonprofit organization dedicated to promoting excellence in language learning, teaching, and research, JALT has a rich tradition of publishing relevant material in its many publications.

Links

- JALT Publications: http://jalt-publications.org
- *JALT Journal*: http://jalt-publications.org/jj
- *The Language Teacher*: http://jalt-publications.org/tlt
- *Conference Proceedings*: http://jalt-publications.org/proceedings

- JALT National: http://jalt.org
- Membership: http://jalt.org/main/membership

JALT
THE JAPAN ASSOCIATION FOR LANGUAGE TEACHING
全国語学教育学会

Articles

The TOEIC IP Test as a Placement Test: Its Potential Formative Value

Christopher Weaver
Toyo University

The use of the Test of English for International Communication Institutional Program (TOEIC IP) at postsecondary institutions in Japan has been increasing over the past decade. The revised TOEIC test provides test-takers and administrators with 3 levels of information: total score, listening and reading sectional scores, as well as the percentage of correct responses achieved on 4 listening subskills and 5 reading skills. Using a hierarchical cluster analysis, this study found significant differences between placement decisions for 1,524 university business students in an English for Specific Purposes program based upon their TOEIC IP scores and their 9 TOEIC subskill scores. The author also discusses some of the diagnostic shortcomings of using a standardized proficiency examination for placement purposes.

日本の高等教育機関におけるTOEIC IPの英語テストの使用は、ここ過去10年増加し続けている。改訂版TOEICテストでは受験者とテスト使用者に合計スコア、リスニング、リーディングという3つのスコアに加え、4つのリスニングサブスキルと5つのリーディングスキルの正答率が提供される。本研究では、階層的クラスター分析を用いて、1,524名の経営学を選考する大学生を対象に調査を行い、TOEIC IPテストのスコアに基づいた英語のクラス分けと、TOEICの 9 つのサブスコアに基づいた英語のクラス分けに有意な差があることを確認した。また、クラス分けの目的で標準化された能力試験を利用する際の診断的な問題点を指摘する。

The use of the Test of English for International Communication Institutional Program (TOEIC IP) continues to increase in education institutions across Japan. The number of TOEIC IP test-takers reached 1,287,456 in 2014 compared to 698,000 test-takers in 2001 (Institute for International Business Communication [IIBC], 2015b). Educational institutions such as high schools, junior colleges, universities, and vocational schools accounted for 45.8% of the TOEIC IP test-takers (589,191 students) in 2014. As a result, teachers and administrators are increasingly facing the need to understand what a TOEIC score means for their students and how it may inform English language programs. Previous empirical investigations have highlighted the importance of curriculum (Hisatsune, 2007), textbook design (Uchibori, Chujo, & Hasegawa, 2006), student study habits (Falout, 2006; Mizumoto & Takeuchi, 2009), and teacher training (Boldt & Ross, 2005) to help facilitate score gains on the TOEIC test. There is, however, a gap in the literature concerning how TOEIC scores can be used to inform placement decisions in a postsecondary EFL program. Of special interest in this paper is the potential for TOEIC IP scores to provide formative feedback to test-takers and administrators.

Placement Examinations

The primary purpose of a placement examination is to create student groups of relatively homogeneous language abilities (Brown, 1996). Typically there are two types of placement examination used to fulfill that purpose (Wall, Clapham, & Alderson, 1994). One type has a preachievement orientation such as the English Placement Test at the University of Illinois at Urbana-Champaign that reflects the academic demands of the courses offered at the university. The other type is standardized tests such as TOEFL, TOEIC, and IELTS that have a general proficiency orientation with no direct relationship to the content of the courses in which test-takers are placed. The use of standardized proficiency examinations has become more common at postsecondary institutions because of the need to evaluate an increasing number of students applying from overseas (Kokhan, 2012; Mullen, 2009). Standardized proficiency examinations are also typically a convenient, cost effective, and rapid means of placing test-takers on the same score scale, which in turn allows relatively easy comparisons and placement of students. Despite these advantages, an increasing number of researchers have become critical of the disconnect between the communicative competence required in different academic settings and the scores achieved on standardized proficiency tests (Chapman & Newfields, 2008; Fox, 2009; Kokhan, 2013).

Central to the argument against the use of standardized proficiency tests for placement purposes is that tests are usually designed to assess a range of knowledge, skills, and abilities within set target domains of language use by a target population (Fulcher & Davidson, 2009). In the case of the TOEIC test, the intention of the test is to measure the everyday English skills of people working in an international environment (Educational Testing Service, 2013). In the case of using the TOEIC test in postsecondary institutions, test-takers may have limited work experience and thus be unfamiliar with the communicative situations featured in this test. Moreover, students' academic interests and ultimate career goals might be quite distant from the world of international business.

The development of the TOEIC test is another cause for concern. The validation process used to evaluate the design of the TOEIC test and the interpretation of its test scores has largely been done in the context of developing a measure of general language proficiency (In'nami & Koizumi, 2012; Tannenbaum & Wylie, 2008) and not a placement examination. As a result, the use of the TOEIC test for a purpose that was not intended is problematic. However, in certain contexts such as in a faculty of business administration or in other postsecondary programs, a test may not only be a practical placement examination but also a valid means of placing university students into an English for Specific Purposes program (IIBC, 2015a). It must be remembered that validity is not a property of the test, but rather concerns the meaning of test scores and the implications for action based upon the interpretation of test scores (Cronbach, 1971; Messick, 1996). From this perspective, placement examination scores should provide test-takers and administrators with diagnostic information about test-takers' strengths and weaknesses in order to group students of similar ability together so that they may receive appropriate materials and instruction (Green & Weir, 2004). This approach to placement examinations, however, relies upon a clear understanding of what a test score means.

Test Scores and Interpretations of Test Scores

A test score describes the interaction between test-takers' performance on the items on a test and the types of knowledge that these items are thought to assess. A test score can be norm-referenced by specifying the relative rankings of the test-takers; it can be criterion-referenced by providing a description of the tasks that test-takers can or cannot perform; or it can be both. The TOEIC test is an attempt to provide test-takers and administrators with information from both of these perspectives. It should be remembered,

however, that any test score includes measurement error depending upon the quality of the test. For example, the standard error of measurement for the TOEIC listening and reading sections is 25 scaled points. Thus if a test-taker has a scaled score of 300 on the listening section of the TOEIC test, 68% of the time this individuals' true score will vary between 275 and 325 (Educational Testing Service, 2013). The standard error of measurement of the TOEIC test may also differ between the listening and reading sections of the test as well as between different populations of test-takers (Zhang, 2006).

Norm-Referenced Information From TOEIC Scores

The official score certificate of the TOEIC test provides test-takers with a total score ranging from 10 to 990, which is the summation of two scaled subscores ranging from 5 to 495 for the listening and reading sections of the test (Educational Testing Service, 2013). Takers of the TOEIC Secure Program (TOEIC SP) test also receive a percentile ranking indicating the percentage of global TOEIC SP test-takers who have a lower scaled score than theirs on the listening and reading sections of the TOEIC test (Educational Testing Service, 2012). This information, however, is not available for TOEIC Bridge and TOEIC IP test-takers (IIBC, 2015b).

Test-takers and administrators need to consult the TOEIC data and analysis report issued annually by the IIBC to understand the relative ranking of test-takers, apart from percentile rankings. For example, in 2014, university students in Japan scored an average of 564 on the TOEIC SP test and 440 on the TOEIC IP test (IIBC, 2015b). Test-takers and administrators can also compare scores against others in the same field of study. In 2014, university students studying commerce, economics, or finance had an average TOEIC IP score of 410 in their first year, 425 in their second year, 476 in their third year, and 516 in their fourth year of study. This seemingly upward trend of TOEIC IP scores, however, may be an artifact of test-taker selection. In 2014, 33,337 first-year students, 18,749 second-year students, 8,643 third-year students, and 2,343 fourth-year students took the TOEIC IP test. The decrease in the number of test-takers each year means that making norm-referenced inferences using TOEIC scores should be carried out with due caution.

Criterion-Referenced Information From TOEIC Scores

In an attempt to help test-takers and administrators interpret TOEIC scores, the Educational Testing Service, the Chauncey Group, and the IIBC undertook a number of initiatives to provide criterion-referenced information for TOEIC scores. TOEIC representatives in Japan developed a 5-level proficiency letter-scale (ranging from A to E) based upon anecdotal information and informal observations on the relationship between TOEIC scores and general English language proficiency. Test-takers and administrators can also consult the score descriptors for the listening and the reading sections to learn about the typical strengths and weaknesses test-takers have. For example, a test-taker with a TOEIC listening score of 250 can typically make simple inferences based on a limited amount of text but cannot make inferences requiring paraphrasing or connecting information (Educational Testing Service, 2007).

To further clarify the relationship between TOEIC scores and the likelihood of being able to perform specific tasks in English, the Educational Testing Service (2000) introduced the *TOEIC Can-Do Guide*. Based upon a correlation analysis of TOEIC IP scores and self-reports of 8,601 Japanese employees (Tannenbaum, Rosenfeld, & Breyer, 1997), the *TOEIC Can-Do Guide* provides a list of different daily life and basic job activities that can be performed at three levels of proficiency: can do, can do with some difficulty, and cannot do. For example, test-takers with a TOEIC reading score ranging from 230 to 350 are thought to be able to understand the type of store or the type of service offered by reading the storefront and to read and understand an agenda for a meeting with some difficulty, but they are thought not to be able to identify differing opinions that opposing party politicians give in two newspaper interviews. Following a similar methodological approach, Ito, Kawaguchi, and Ota (2005) developed a can-do list focusing upon 65 job-related tasks occurring in seven different communicative situations. Based upon the responses of 8,386 Japanese company employees, TOEIC scores ranging from 400 to 495 were reported as the point where people began feeling comfortable about using English to do job-related tasks such as reading a manual written in English about office equipment or sending an email to a company to complain about a product.

The Redesigned TOEIC Test and Scores

One of the main difficulties in interpreting TOEIC scores is that test-takers and administrators need to consult outside resources that attempt to link TOEIC scores to certain communicative competencies, the Language Profi-

ciency Interview test (Wilson, 1989), or the Common European Framework of Reference for languages (Tannenbaum & Wylie, 2008). As a response to this difficulty, one of the goals for the redesigned TOEIC test was to provide more specific information about test-takers' abilities based upon their test performance (Liao, 2010). To do so, the designers of the revised TOEIC test used the Evidence-Centered Design (ECD) method. ECD is an evidentiary reasoning approach to assessment design that (a) identifies the important domains of knowledge to be assessed and ascertains how this knowledge is acquired and used; (b) establishes a chain of reasoning between what individuals say and do in assessments and the inferences that can be made about what test scores mean in terms of the abilities that individuals currently possess and what should be done next; and (c) ensures that test design reflects the purpose of the assessment while taking into consideration the constraints, resources, and conditions of use (Mislevy, Almond, & Lukas, 2003). These three premises guided the new item prototypes and pilot testing of the revised TOEIC test to ensure a sufficient distribution of items that could provide reliable support for claims about test-taker abilities (Schedl, 2010). The abilities measured in the TOEIC test (shown in Table 1) are articulated in four listening subskills and five reading subskills.

Table 1. TOEIC Test Subskills

Listening subskills	
L1	Can infer gist, purpose, and basic context based on information that is explicitly stated in short spoken texts.
L2	Can infer gist, purpose, and basic context based on information that is explicitly stated in extended texts.
L3	Can understand details in short spoken passages.
L4	Can understand details in extended spoken texts.
Reading subskills	
R1	Can make inferences based on information in written texts.
R2	Can locate and understand specific information in written texts.
R3	Can connect information across multiple sentences in a single written text and across texts.
R4	Can understand vocabulary in written texts.
R5	Can understand grammar in written texts.

(Educational Testing Service, 2008)

These ECD-based claims about test-takers' listening and reading abilities give rise to a situation in which the TOEIC subskill scores may potentially provide test administrators with more detailed information that could enable them to make more nuanced placement decisions in a postsecondary EFL education setting. From the perspective of formative assessment, teachers and program administrators may be able to identify specific gaps in communicative competence that are exposed in test-takers' performance on the test (Long & Norris, 2000). This possibility warrants an investigation into the use of TOEIC scores and TOEIC subskill scores for placement purposes.

The following research questions guided this investigation:

RQ1. To what extent do test-takers' TOEIC IP scores relate to the average percentage of correct responses they achieved on the nine TOEIC subskills?

RQ2. To what extent can test-takers be grouped together according to the percentage of correct answers they achieved on the nine TOEIC subskills?

RQ3. To what extent do test-taker groupings based upon TOEIC IP scores differ from test-taker groupings based upon the percentage of correct answers achieved on the nine TOEIC IP subskills?

The first research question examines the degree to which the nine TOEIC subskills provide unique test-taker information compared to the total TOEIC IP score. If the subskills are highly related to TOEIC IP score or the sectional scores, then there is little reason to spend extra time and resources, which are typically in short supply in placement situations, to examine the test-takers' performances on the nine TOEIC subskills. However if the nine TOEIC subskills provide unique information about the test-takers, is it then possible to create student groups more sensitive to test-takers' abilities? This is the focus of the second research question. The final research question examines the practical effect of student placements based upon TOEIC scores and the test-takers' performance on the nine TOEIC subskills. In other words, does more refined test-taker information really make a difference when it comes to actually placing them into a postsecondary EFL program?

Participants

The test-takers in this study were 1,524 university students (883 men and 641 women) studying in the faculty of business administration at a private university located in Tokyo during the 2012-13 academic year. This population represents a fraction of the TOEIC IP test-takers in Japan and this needs

to be kept in mind when considering the generalizability of the findings of this study. The test data originated from a TOEIC IP test administered at the university by a group of trained proctors from the IIBC in the fall semester. The TOEIC IP test was a requirement for the students to receive credit for a required English course. The results of the TOEIC IP test were also used to place the students into their subsequent English courses.

Analysis

To answer the first research question, Pearson correlation coefficients between the TOEIC IP score, the listening section test scores, the reading section test scores, and the percentage of correct responses achieved on the nine TOEIC subskills were examined. The magnitude of these correlations was also evaluated using the reduction in uncertainty (RiU) index (Dorans, 2004). This statistic measures the amount of certainty that the total test score and the scores from the listening and the reading sections are similar to the percentage of correct responses achieved on the nine TOEIC subskills. Dorans suggested that a 50% RiU in one observed score from another observed score is needed to confirm a linkage between the test scores.

To answer the second research question, a hierarchical cluster analysis was used to classify the test-takers into groups based upon the percentage of correct answers they achieved on the nine subskills. Similar to a factor analysis, a cluster analysis examines the interrelationships between the variables; however, there are no preconceived ideas about the composition of the groups in a cluster analysis. The formation of the clusters is informed by the analysis of the data (Everitt, Landau, & Leese, 2001) and thus can help identify individual differences that exist between language learners (Skehan, 1989). Once the clusters were identified, a series of univariate one-way ANOVAs were used to determine which subskills significantly differed between the groups of test-takers. To offset the chances of a Type I error (Simes, 1986), a Bonferroni adjustment was used to set a p value of 0.005 for this study.

To answer the third research question, Cramer's V was used to determine the degree of agreement between the test-taker groupings based upon their total TOEIC IP scores and the test-taker groupings based upon the cluster analysis of the nine subskills. This nonparametric statistic measures the degree of strength between categorical variables that have more than two categories (Sheskin, 2007). A Cramer's V of 0 reflects complete independence between the categorical variables, whereas a Cramer's V of 1 indicates

a complete association or dependence between the variables.

Results

Table 2 shows that the mean TOEIC IP score for the test-takers in this study was 352.32 with a standard deviation of 91.91. These test-takers had a higher level of listening proficiency (M = 205.89, SD = 53.59) compared to their reading proficiency (M = 146, SD = 48.24). These test-takers had lower test scores compared to other university students who took the TOEIC IP test during the same time period in Japan (i.e., a mean score of 433 for the TOEIC IP test, 245 for the listening section, and 188 for the reading section), according to the 2012 TOEIC analysis report (IIBC, 2013).

Table 2. Test-takers' TOEIC IP Scores, Listening and Reading Section Scores, and the Percentage of Correct Responses Achieved on the Nine TOEIC Subskills (N = 1,524)

	M	*SD*	Min	Max	Kurtosis	*SEM*
TOEIC IP	352.32	91.91	150	765	0.53	0.13
Listening section	205.89	53.59	60	430	0.29	0.13
Reading section	146.43	48.24	30	365	0.71	0.13
Listening subskill L1	58.05	12.59	5	95	0.08	0.13
Listening subskill L2	46.46	14.11	8	92	-0.22	0.13
Listening subskill L3	60.35	14.93	5	95	0.05	0.13
Listening subskill L4	42.91	12.31	11	89	0.04	0.13
Reading subskill R1	31.98	13.78	0	88	0.32	0.13
Reading subskill R2	40.19	14.61	0	90	-0.21	0.13
Reading subskill R3	30.60	12.10	0	88	0.73	0.13
Reading subskill R4	34.74	10.59	7	72	-0.02	0.13
Reading subskill R5	48.70	13.95	4	92	-0.23	0.13

Note. Test scores for the TOEIC IP test, the listening section, and the reading section are reported as scaled points, whereas scores for the TOEIC subskills are reported as a percentage of correct answers. *SEM* = standard error of measurement.

In terms of the TOEIC listening subskills, these test-takers had the highest

percentage of correct answers (60.35%) on items designed to assess the understanding of details in a short passage (L3) and the lowest percentage of correct answers (42.91%) on test items designed to assess the understanding of details in extended spoken texts (L4). For the TOEIC reading subskills, the test-takers had the highest percentage of correct answers (48.7%) on items designed to assess the understanding of grammar in written texts (R5) and the lowest percentage of correct answers (30.6%) on test items designed to assess the ability to connect information across multiple sentences in a single written text and across texts (R3).

Tables 3 and 4 show the Pearson correlation coefficients and magnitudes for the TOEIC IP score, the listening and reading section scores, and the percentage of correct responses achieved on the nine TOEIC subskills. The correlations between the total test score and the sectional scores are relatively high, with the highest correlation being between the total test score and the listening section score (.91). The correlations between the percentage of correct responses achieved on the nine TOEIC subskills, the TOEIC IP score, and the sectional scores for listening and reading are small to moderate. The listening subskill showing the highest correlation with the total test score is L4 (can understand details in extended spoken texts) at .77 with an RiU of 36% between these two scores. The reading subskills showing the highest correlations with the total test score are R5 (can understand grammar in written texts) at .68 with an RiU of 26% and R2 (can locate and understand specific information in written texts) at .67 with an RiU of 26%.

A hierarchical cluster analysis using Ward's minimum variance method with the squared Euclidean distance technique (Szekely & Rizzo, 2005) was run on the percentage of correct answers achieved by the test-takers on the nine TOEIC subskills. In the case of this analysis, a four-cluster solution was the point at which subsequent clusters did not significantly add to the process of distinguishing the test-takers from each other (Burns & Burns, 2008). Figure 1 shows how the four clusters of test-takers performed on the nine subskills.

Table 3. Correlations Among TOEIC IP (T) Score, Listening (L) and Reading (R) Section Scores, and the Percentage of Correct Responses Achieved on the Nine TOEIC Subskills

	L	R	L1	L2	L3	L4	R1	R2	R3	R4	R5
T	.91	.89	.63	.73	.67	.77	.52	.67	.53	.59	.68
L		.63	.68	.80	.74	.85	.35	.49	.35	.38	.50
R			.45	.51	.45	.52	.59	.73	.63	.69	.74
L1				.39	.45	.43	.26	.32	.24	.31	.37
L2					.46	.58	.28	.42	.30	.31	.39
L3						.47	.23	.34	.22	.26	.40
L4							.30	.42	.30	.31	.39
R1								.42	.47	.21	.23
R2									.40	.28	.36
R3										.33	.36
R4											.44

Table 4. Reduction in Uncertainty (RiU) Among the Total TOEIC IP (T) Score, the Listening (L) and Reading (R) Section Scores, and the Percentage of Correct Responses Achieved on the Nine TOEIC Subskills

	T	L	R
L1	22%	26%	11%
L2	32%	39%	14%
L3	25%	33%	10%
L4	36%	47%	15%
R1	14%	6%	20%
R2	26%	13%	32%
R3	15%	6%	22%
R4	19%	8%	28%
R5	26%	13%	33%

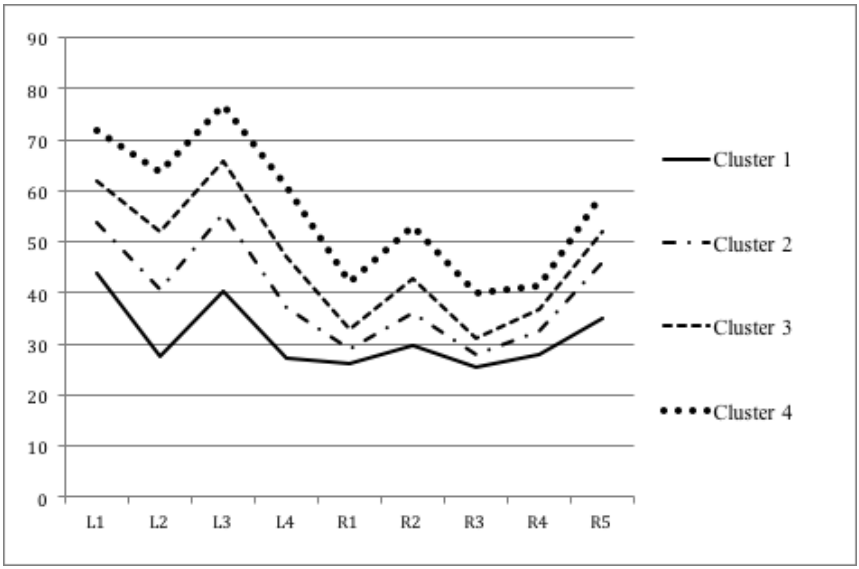

***Figure 1*. The percentage of correct responses on the nine TOEIC subskills for the four test-taker clusters.**

In regards to the TOEIC listening subskills, the four clusters of test-takers followed a similar pattern. They had a higher percentage of correct responses on items designed to assess their ability to infer gist, purpose, and basic context based on information that is explicitly stated in short spoken texts (L1) and their ability to understand details in short spoken passages (L3). In contrast, test items featuring extended spoken texts (subskills L2 and L4) were more challenging.

For the five TOEIC reading subskills, the four clusters of test-takers had some distinctive characteristics. The proficiency level of Cluster 1 was relatively flat across the five reading subskills compared to the other clusters of test-takers. Cluster 2 shared Cluster 1's difficulty on items designed to assess the ability to make inferences based on information in written texts (R1) and to assess the ability to connect information across multiple sentences in a single written text and across texts (R3). Clusters 3 and 4 followed a similar pattern with higher rates of success on items designed to assess the ability to locate and understand specific information in written texts (R2) and items assessing the ability to understand grammar in written texts (R5).

Nine univariate one-way ANOVAs found that the four clusters significantly differed from each other in terms of the percentage of correct responses they achieved on the nine subskills. The right column of Table 5 shows that only two out of the 45 possible post hoc pairwise comparisons were not significantly different (i.e., Clusters 1 and 2 are not significantly different for the TOEIC reading subskills R1 and R3). The results of the Tukey post hoc test comparisons were reconfirmed with the Games-Howell procedure (Wilcox, 1987) because there was an unequal number of test-takers in each cluster.

Table 5. The Percentage of Correct Answers Achieved on the TOEIC Subskills for the Four Clusters

TOEIC subskills	Cluster 1 (n = 187)	Cluster 2 (n = 589)	Cluster 3 (n = 520)	Cluster 4 (n = 228)	Post hoc comparison of clusters
L1	44.65 (11.51)	54.67 (10.36)	61.43 (9.96)	70.04 (10.08)	1-2, 1-3, 1-4, 2-3, 2-4, 3-4
L2	30.14 (10.37)	42.71 (11.42)	49.77 (11.07)	62.02 (10.21)	1-2, 1-3, 1-4, 2-3, 2-4, 3-4
L3	39.49 (12.34)	57.32 (11.78)	64.89 (10.38)	74.96 (11.01)	1-2, 1-3, 1-4, 2-3, 2-4, 3-4
L4	30.26 (8.09)	38.87 (9.62)	45.53 (9.31)	57.7 (10.75)	1-2, 1-3, 1-4, 2-3, 2-4, 3-4
R1	25.61 (8.81)	25.49 (11.14)	34.60 (10.89)	47.97 (14.18)	1-3, 1-4, 2-3, 2-4, 3-4
R2	28.24 (10.01)	33.52 (11.26)	45.28 (12.12)	55.57 (12.46)	1-2, 1-3, 1-4, 2-3, 2-4, 3-4
R3	24.66 (11.25)	25.31 (9.03)	33.20 (9.88)	43.22 (12.91)	1-3, 1-4, 2-3, 2-4, 3-4
R4	26.57 (8.04)	32.33 (8.86)	36.60 (10.10)	43.42 (10.43)	1-2, 1-3, 1-4, 2-3, 2-4, 3-4
R5	31.17 (10.27)	45.22 (10.65)	53.18 (11.27)	61.48 (11.40)	1-2, 1-3, 1-4, 2-3, 2-4, 3-4

Note. *M* (*SD)*; all post hoc tests assessing differences between student clusters were set to $p < .005$.

The degree of agreement between test-taker groupings based upon their TOEIC IP scores and the percentage of correct answers they achieved on the nine TOEIC subskills was large, a Cramer's V of .63 (Cohen, 1988). Table 6 shows the crosstabulation table for these two categorical groupings. Test-takers with the lowest test scores (in the 100s) and the highest scores (in the 500s and 600s) generally grouped together in Cluster 1 and Cluster 4 respectively. The majority of test-takers with test scores ranging from the 200s to the 400s belonged to Clusters 1, 2, or 3.

Table 6. Crosstabulation Table of TOEIC IP Scores and TOEIC Subskill Clusters

TOEIC IP score	TOEIC subskill clusters				
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
600s	0	0	0	20	20
500s	0	**0**	**2**	73	75
400s	0	9	187	133	329
300s	**1**	345	323	**2**	671
200s	136	233	**8**	0	377
100s	50	**2**	0	0	52
Total	187	589	520	228	1,524

Note. Bold numbers denote test-takers who were outliers in their cluster membership.

To check that clusters did not simply reflect a score difference of 50 scaled TOEIC points, the cluster memberships of test-takers with scores ranging from 300 to 345 and of those of test-takers with scores ranging from 350 to 395 were examined. In Cluster 2, there were 249 test-takers with scores between 300 and 345 and 96 test-takers with scores between 350 and 395. A similar split occurred in Cluster 3: 88 test-takers had scores between 300 and 345; 235 test-takers had scores between 350 and 395. In sum, about 28% of the test-takers belonged to a different cluster than did other test-takers with a similar TOEIC IP score.

Table 6 also shows that 15 test-takers were outliers in their cluster memberships (highlighted in bold). In Cluster 1, there was one test-taker with a score in the 300s who correctly answered 30% of the items for all five

reading subskills. Cluster 2 had two test-takers who had a higher percentage of correct answers on items designed to assess the ability to understand vocabulary (R4) and grammar (R5) in written texts. In Cluster 3, there were eight test-takers with a score in the 200s who had significantly higher scores and two test-takers with a score in the 500s who had significantly lower scores on the five reading subskills. Cluster 4 had one test-taker with a score in the 300s who had significantly higher scores on the four listening subskills and another test-taker who had significantly higher scores on the five reading subskills.

Discussion

Understanding the meaning of test scores is central to well-informed placement decisions. In the case of the TOEIC IP test, interpretations of the total test score and the listening and reading section test scores can be clarified with the percentage of correct answers test-takers achieved on the four listening and five reading subskills. In this study, the correlations between the test-takers' TOEIC IP scores and the percentage of correct responses they achieved on the nine TOEIC subskills were all moderately correlated. Yet, the magnitude of these correlations is below Dorans's 50% recommendation for the RiU index. These low percentages are understandable considering that the nine subskills are used together to calculate the test-takers' subsection and total test scores. However, the subskills that have larger percentages on the uncertainty index highlight competencies that may benefit from subsequent EFL instruction. For example, the stronger link between the listening section test scores, listening subskill L2 (the ability to infer gist, purpose, and basic context based on information that is explicitly stated in extended texts), and listening subskill L4 (the ability to understand details in extended spoken texts) suggests the importance of providing students with the opportunity to hear and act upon extended spoken texts within the classroom.

Beyond correlation analysis, a hierarchical cluster analysis of the nine TOEIC subskills provides test administrators with a graphical representation of the test-takers' strengths and weaknesses (see Figure 1). In this study, test-takers had the greatest difficulty with items that required them to comprehend extended spoken texts and make inferences or connect information in written texts. This type of information can in turn help define cut-scores and allow for more nuanced placement decisions (Powers & Powers, 2014). Table 6 shows two possible ways in which the test-takers can be divided into their classes. The first way is to simply use the total TOEIC IP

score or the listening and reading subsection test scores. The result would be a continuum of classes that range from test-takers who have a score in the 100s to test-takers with score in the 600s. This vertical approach, however, ignores that there are different ways to reach the same TOEIC IP score. Administrators and teachers as a result have limited information that could inform curriculum design and implementation.

The second way is take into account the listening and reading subskills that comprise the TOEIC test. Reading Table 6 horizontally reveals the group memberships that exist within the vertical continuum of TOEIC IP scores. The test-takers with the lowest scores (i.e., in the 100s) and the highest scores (i.e., in the 500s and 600s) predominately belonged to Cluster 1 and Cluster 4 respectively. In contrast, test-takers with scores ranging from the 200s to the 500s were split between two or three different clusters of students. Using this information, initial placement decisions can be made according to the total TOEIC IP scores and final placements can take into consideration the test-takers' level of success on the nine subskills. Administrators and teachers would then have the needed information to select materials to address specific listening and reading skills. There is a meaningful difference between (a) informing a teacher that a class has a mean score of 187.17 for the listening section and 121.83 for the reading section of the TOEIC IP test and (b) telling that teacher the class also belongs to Cluster 2, which generally has difficulty understanding details in extended spoken texts (L4), making inferences based on information in written texts (R1), and connecting information across multiple sentences in a single written text and across texts (R4). In short, the criterion-referenced information included in the nine TOEIC subskills can help test-takers and administrators gain a better understanding of what a TOEIC score actually means.

The analysis of the nine TOEIC subskills has the additional benefit of potentially identifying test-takers who have unique strengths and weaknesses. In this study, there were 15 students who had significantly different skill sets when compared to test-takers with similar test scores. Although this group of outlying students is only 1% of the test-takers, their needs should not be overlooked. Ideally placement decisions should be responsive to individual needs and not group norms.

There are, however, limits to the diagnostic information that the TOEIC IP test can provide. The nine TOEIC subskills are not as fine-grained as the subskills used in other standardized proficiency tests (see, e.g., Kim, 2014). The TOEFL test, for example, divides the TOEIC reading subskill R4 (understanding vocabulary in a written text) into deducing word meaning from context

or without context (Jang, 2005). These more refined subskills can in turn help teachers and materials designers to develop lesson plans to address specific test-taker strengths and weaknesses. The TOEIC IP test also does not assess test-takers' level of productive English competence. As a result, score interpretation and placement decisions into four-skills language programs can be problematic when based on TOEIC IP scores (Mullen, 2009). A possible remedy might be administering the speaking and writing sections of the TOEIC test. However, the additional cost and resources required to test large groups of students may threaten the practicality of using the TOEIC IP test for placement purposes.

Although placement examinations are not necessarily high-stakes tests, they can have significant consequences on test-takers' chances of success in a program after they have been placed. As a result, there is a need for future investigations into the trustworthiness of inferences made from standardized proficiency test scores and placement decisions made based upon these scores. In addition, placement examinations that not only assess linguistic ability but also take into consideration test-takers' domains of language use (see Thompson, 2015) or their willingness to use their L2 require further investigation. Future research might also examine what type of information stakeholders such as teachers consider informative in the placement decision process as well as in the subsequent implementation of the language program (see Fox, 2009).

Despite a number of concerns surrounding the use of standardized proficiency examinations for placement purposes, the ever-increasing use of the TOEIC IP test at educational institutions in Japan gives rise to the need to carefully consider the potential formative value of TOEIC IP scores. This test provides test-takers and administrators with three levels of information: the total score, the listening and the reading section scores, and the percentage of correct responses for four listening and five reading TOEIC subskills. These scores are interrelated, but they provide unique vantage points that can be used to identify groups of test-takers based upon their strengths and weaknesses, which in turn have the potential of facilitating nuanced placements and more targeted language instruction.

Christopher Weaver is an associate professor in the Faculty of Business Administration at Toyo University, Tokyo, Japan. His research focuses upon assessment, task-based instruction, and willingness to communicate, with the purpose of facilitating opportunities for second language development.

References

Boldt, R., & Ross, S. (2005). *Language proficiency gain on the Test of English for International Communication: Meta-analysis of Japanese and Korean corporate language programs.* TOEIC Research Report. Retrieved from http://www.toeic.or.jp/library/toeic_data/toeic/data/pdf/boldt_ross2005.pdf

Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice-Hall Regents.

Burns, R., & Burns, R. (2008). *Business research methods and statistics using SPSS*. London, UK: Sage.

Chapman, M., & Newfields, T. (2008). The "new" TOEIC. *Shiken: JALT Testing & Evaluation SIG Newsletter, 12*(2), 32-37.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cronbach, L. (1971). Test validation. In R. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

Dorans, N. (2004). Equating, concordance, and expectation. *Applied Psychological Measurement, 28,* 227-246. http://dx.doi.org/10.1177/0146621604265031

Educational Testing Service. (2000). *TOEIC can-do guide: Linking TOEIC scores to activities performed using English*. Retrieved from http://www.ets.org/Media/Research/pdf/TOEIC_CAN_DO.pdf

Educational Testing Service. (2007). *TOEIC listening score descriptors. TOEIC reading score descriptors.* Retrieved from https://www.ets.org/Media/Tests/TOEIC/pdf/TOEIC_LR_Score_Desc.pdf

Educational Testing Service. (2008). *TOEIC listening and reading: Official score report*. Retrieved from https://www.etsglobal.org/Pl/Pol/content/download/1461/25305/version/1/file/TOEIC-score-report.pdf

Educational Testing Service. (2012). *TOEIC listening & reading percentile rank*. Retrieved from https://www.ets.org/s/toeic/pdf/toeic_listening_and_reading_percentile_rank.pdf

Educational Testing Service. (2013). *TOEIC user guide: Listening and reading*. Retrieved from http://www.ets.org/Media/Tests/Test_of_English_for_International_Communication/TOEIC_User_Gd.pdf

Everitt, B. S., Landau, S., & Leese, M. (2001). *Cluster analysis* (4th ed.). London, UK: Arnold.

Falout, J. (2006). Japanese college laboratory of science majors preparing for the TOEIC. In K. Bradford-Watts (Ed.), *JALT2005 Conference Proceedings* (pp. 1140-1151). Tokyo: JALT.

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,* 26-42. http://dx.doi.org/10.1016/j.jeap.2008.12.004

Fulcher, G., & Davidson, F. (2009). Test architure, test retrofit. *Language Testing, 26,* 123-144. http://dx.doi.org/10.1177/0265532208097339

Green, A., & Weir, C. (2004). Can placement tests inform instructional decisions? *Language Testing, 21,* 467-494. http://dx.doi.org/10.1191/0265532204lt293oa

Hisatsune, A. (2007). Meet the challenges: Empowering TOEIC students. In K. Bradford-Watts (Ed.), *JALT2006 Conference Proceedings* (pp. 315-331). Tokyo: JALT.

In'nami, Y., & Koizumi, R. (2012). Factor structure of the revised TOEIC test: A multiple-sample analysis. *Language Testing, 29,* 131-152. http://dx.doi.org/10.1177/0265532211413444

Institute for International Business Communication. (2013). *TOEIC test data & analysis 2012*. Retrieved from http://www.toeic.or.jp/library/toeic_data/toeic_en/pdf/data/TOEIC_Program_DAA2012.pdf

Institute for International Business Communication. (2015a). *TOEIC Newsletter No. 125*. Retrieved from http://www.toeic.or.jp/library/toeic_data/toeic_en/pdf/newsletter/newsletterdigest125.pdf

Institute for International Business Communication. (2015b). *TOEIC test data & analysis 2014*. Retrieved from http://www.toeic.or.jp/library/toeic_data/toeic_en/pdf/data/TOEIC_Program_DAA.pdf

Ito, T., Kawaguchi, K., & Ohta, R. (2005). *A study of the relationship between TOEIC scores and functional job performance: Self-assessment of foreign language proficiency.* Retrieved from http://www.toeic.or.jp/library/toeic_data/toeic_en/pdf/newsletter/1_E.pdf

Jang, E. (2005). *A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (AAT 3182288)

Kim, A. (2014). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing*, *32,* 227-258. http://dx.doi.org/10.1177/0265532214558457

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,* 291-308. http://dx.doi.org/10.1177/0265532211429403

Kokhan, K. (2013). An argument against using standardized test scores for placement of international undergraduate students in English as a second language (ESL) courses. *Language Testing, 30*, 467-489. http://dx.doi.org/10.1177/0265532213475782

Liao, C. (2010). *TOEIC listening and reading test scale anchoring study*. Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/TC-10-05.pdf

Long, M., & Norris, J. (2000). Task-based language teaching and assessment. In M. Byram (Ed.), *Encyclopoedia of language teaching* (pp. 597-603). London, UK: Routledge.

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

Mislevy, R., Almond, J., & Lukas, J. (2003). *A brief introduction to evidence-centered design.* Princeton, NJ: Educational Testing Service. Retrieved from https://www.ets.org/Media/Research/pdf/RR-03-16.pdf

Mizumoto, A., & Takeuchi, O. (2009). A closer look at the relationship between vocabulary learning strategies and the TOEIC scores*. TOEIC Research Report, 4,* 1-34.

Mullen, A. (2009). *The impact of using a proficiency test as a placement tool: The case of the Test of English for International Communication (TOEIC)* (Doctoral dissertation). Retrieved from www.theses.ulaval.ca/2009/26672/26672.pdf

Powers, D., & Powers, A. (2014). The incremental contribution of TOEIC Listening, Reading, Speaking, and Writing tests to predicting performance on real-life English language tasks. *Language Testing, 32*, 151-167. http://dx.doi.org/10.1177/0265532214551855

Schedl, M. (2010). *Background and goals of the TOEIC Listening and Reading test redesign project*. Princeton, NJ: Educational Testing Service. Retrieved from https://www.ets.org/Media/Research/pdf/TC-10-02.pdf

Sheskin, D. (2007). *Handbook of parametric and nonparametric statistical procedures* (4th ed.). Boca Raton, FL: Chapman & Hall/CRC.

Simes, R. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika, 73,* 751-754. http://dx.doi.org/10.1093/biomet/73.3.751

Skehan, P. (1989). *Individual differences in second-language learning*. New York, NY: Routledge.

Szekely, G., & Rizzo, M. (2005). Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method. *Journal of Classification, 22,* 151-183. http://dx.doi.org/10.1007/s00357-005-0012-9

Tannenbaum, R., Rosenfeld, M., & Breyer, F. (1997). *Linking TOEIC score to self-assessments of English-language abilities: A study of score interpretation*. Unpublished manuscript.

Tannenbaum, R., & Wylie, E. (2008). *Linking English-language scores onto the Common European Framework of Reference: An application of standard-setting methodology.* Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/RR-08-34.pdf

Thompson, G. (2015). Understanding the heritage language student: Proficiency and placement. *Journal of Hispanic Higher Education, 14,* 82-96. http://dx.doi.org/10.1177/1538192714551277

Uchibori, A., Chujo, K., & Hasegawa, S. (2006). Towards better grammar instruction: Bridging the gap between high school textbooks and TOEIC. *The Asian EFL Journal Quarterly, 8,* 228-253. Retrieved from http://asian-efl-journal.com/June_2006_EBook_editions.pdf

Wall, D. , Clapham, C., & Alderson, J. C. (1994). Evaluating a placement test. *Language Testing, 11,* 321-344. http://dx.doi.org/10.1177/026553229401100305

Wilcox, R. (1987). *New statistical procedures for the social sciences: Modern solutions to basic problems*. Hillsdale, NJ: Erlbaum.

Wilson, K. (1989). *Enhancing the interpretation of a norm-referenced second-language test through criterion referencing: A research assessment of experience in the TOEIC testing context.* Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1989.tb00153.x

Zhang, S. (2006). Investigating the relative effects of persons, items, sections, and languages on TOEIC score dependability. *Language Testing, 23,* 351-369. http://dx.doi.org/10.1191/0265532206lt332oa