Kumazawa et al. Language Testing in Asia (2016) 6:2
DOI 10.1186/s40468-015-0023-3

Language Testing in Asia
a SpringerOpen Journal

RESEARCH Open Access

Validity argument for the VELC Test® score interpretations and uses

Takaaki Kumazawa[1*], Tetsuhito Shizuka[2], Masamichi Mochizuki[3] and Atsushi Mizumoto[4]

* Correspondence:
ktakaaki@kanto-gakuin.ac.jpAQ2
[1]Kanto Gakuin University, Odawara, Kanagawa, Japan
Full list of author information is available at the end of the article

Abstract

Background: Placement testing is a crucial issue in Japanese universities. In the majority of language programs, classes are streamed by proficiency levels based on students' placement test score for efficient instruction because university students' proficiency levels vary greatly even in the same program. The Visualizing English Language Competency Test® [VELC Test® (VELC Research Group, 2013)] was designed particularly for making Japanese university students' proficiency and placement decisions.

Methods: This study provides a validity argument for the VELC Test® score interpretations and uses using Kane's (2006) argument-based validity framework when administered to 4407 Japanese university students as a placement test.

Results: Four inferences from observation to decision were adequately made due to the facts that: (a) most of the VELC Test® items were working as placement items (scoring), (b) the VELC Test® ($k = 120$) was reliable with the small amount of error (generalization), (c) test-takers' VELC Test® score could show what they could do with their English (extrapolation), and (d) the VELC Test® could be used to separate test-takers' proficiency into three levels and could be useful for test-takers' further learning (decision).

Conclusions: This study indicated that administrators could make valid Japanese university students' placement decisions with this test.

Keywords: Validity argument, Placement testing, Generalizability theory, Rasch model

Background

Placement tests, both commercial and in-house, are widely adopted in language programs to place students into their appropriate levels (Brown et al. 2004). At Japanese universities where students' English proficiency levels differ to a great degree even within the same institution, most English programs have adopted a level-grouping curriculum where placement tests are administered and the scores are interpreted and used to make placement decisions. The Visualizing English Language Competency Test® [VELC Test® (VELC Research Group, 2013)] was developed for the purpose of estimating test-takers' proficiency levels on vocabulary, grammar, reading, and listening skills, and placing Japanese university students into their appropriate level. Some validation studies on the commercial test (e.g., Shizuka & Mochizuki 2014a) have already been done, but more studies are needed to collect the test validity evidence using a current validity framework.

Validation studies on placement tests (e.g., Wall et al. 1994) have been done using the traditional concept of test validity. The concept has drastically evolved in a few decades. Along with the change, one of the most prominent validity frameworks called an argument-based validity framework has been proposed (e.g., Bachman & Palmer 2010; Chapelle 2008; Kane 2006). Some validation studies for high-stakes tests such as the TOEFL® have been conducted using the framework (e.g., Chapelle et al. 2010) but almost none has been done on placement tests. It is test-developers' responsibility to publicize results obtained from validation studies in order to avoid users' misuse of the tests. Thus, the purpose of this study is to present a validity argument for the VELT Test® score interpretations using Kane's (2006) argument-based validity framework.

Argument-based validity framework

Traditionally, validity is defined as the degree to which a test measures what it claims, or purports to be measuring (Chapelle 2012). To validate a test, language testers have to examine criterion-related, content and/or construct validity in addition to reliability. To examine criterion-related validity, testers compute correlation between test scores that testers try to validate and the ones that have been validated. The coefficient is known as a validity coefficient. Content validity can be scrutinized by having experts judging the extent to which test items are sampled from a target domain. Statistics such as factor analysis is used to reveal what factors can be extracted from test items. Cronbach and Meehl (1955) pointed out the need to go beyond the traditional concept of validity and proposed a more unified concept.

According to the Standards (AERA/APA/NCME 1999), validity refers to "the degree to which evidence and theory support the interpretations of the scores entailed by proposed uses of tests" (p. 9). Thus, validation is the evaluation of test score interpretations and uses by first providing proposed test score interpretations and uses and then arguing for the proposed interpretations and uses based on sound warrants. To validate test score interpretations and uses, several validity frameworks are proposed in the fields of educational measurement (Kane 2006; Messick 1995) and language testing (Bachman 2005; Chapelle 2008) and empirical studies are conducted using the frameworks (e.g., Beglar 2010; Llosa 2008; Pardo-Ballester 2010).

A major advantage of using Kane's (2006) argument-based approach to validity is that it "provides a relatively pragmatic approach to validation" (Kane 2012, p. 15). In Kane's framework (2006), testers make three chains of inference for test score interpretations and one for test score uses. The four inferences are (a) scoring from the observation to the observed score, (b) generalizability from the observed score to the universe score, (c) extrapolation from the universe score to the target score, and (d) decision from the target score to the use (see Fig. 1). The proposed argument that the four bridges of inference are theoretically valid is called an interpretative argument. Then, language testers collect any necessary backings so that testers can provide warrants to argue for the interpretive argument. The statements are called a validity argument. Thus, to validate test score interpretations and uses, testers first provide an interpretive argument that shows that test items work as intended, are a representative of a universe domain, are a representative of a target domain, and the score is used to make sound decisions. Finally, testers propose a validity argument based on backings collected in order to evaluate the interpretive argument presented.

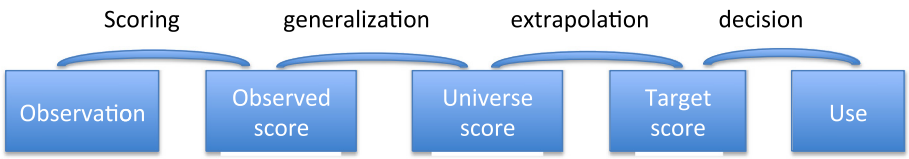

Fig. 1 Four inferences in Kane's argument-based approach to validation (adapted from Kane, Crooks, & Cohen 1999)

Placement testing

In the field of language testing, placement testing is a major issue (e.g., Bachman 1990; Brown 2005). In order for students to receive appropriate academic support and optimal levels of instruction, language programs have generally adopted norm-referenced tests to place students based on their language ability. As a result of successful placement, students are unlikely to experience academic boredom, frustration, and in the worst case, failure of the course. Thus, placement is a crucial element in most language programs.

A survey showed that out of 169 American university language programs, 122 programs (72 %) administer such tests (Brown et al. 2004). While 87 programs (51 %) adopted commercial tests, 54 programs (32 %) developed their in-house placement tests and 19 (11 %) used them both. The major reason for adopting a commercial test is that tests such as the Advanced Placement® test is widely used in other programs and the score is consistent. The major argument for developing a test inside an institution is that the test content can be linked to the program's course objectives. Whether to adopt commercial tests or develop in-house placement tests is a debatable issue.

Papageorgiou and Cho (2013) discussed the use of the TOEFL Junior™ Standard test for placing ESL students in secondary education. The test was given to 92 ESL students in two schools and the teachers also judged what level they had to be placed. The strong correlation between the levels identified by the test score and the judgment supported the use of the test for placement purposes. Kokhan (2013), however, presented an argument against the use of commercial tests for placement. She investigated the use of the SAT, ACT, and TOEFL iBT® as a replacement for an in-house developed placement test for making ESL university students' placement decisions. She concluded that these commercial tests might not be appropriate for placement purposes because their testing purposes (i.e., proficiency, and placement) were different.

Studies on in-house placements generally support the use of locally made tests. Brown (1989) first argued that placement tests had to not only create variance in the score but also be related to content and skills covered in a program. Then, he demonstrated how an in-house placement test could be revised to link the test content to the program content by taking an item-analysis strategy. Sasaki (1991) compared three methods of item differential functioning (DIF) and claimed that DIF could be used to detect placement items that a certain group did not master so that a remediation could be provided. Fulcher (1997) examined validity of an in-house placement test and provided evidence for the reliability, construct, concurrent, and content validity. Two studies investigated dimensionality of in-house placement tests (Blais & Laurier 1995; Green & Weir 2004) and found that the test reliability being high and unidimensionality being maintained. Green and Weir (2004), however, claimed that their in-house placement test was not useful for identifying test-takers' mastery of grammar. Jamieson,

Wang, and Church (2013) compared the use of both in-house and commercial placement speaking tests in terms of curricular coverage, statistical distributions, and practicality. Because of the high practicality and the low cost, they argued for the use of the in-house placement test.

In the Japanese context, most studies are conducted on commercially produced placement tests. Culligan and Gorsuch (1999) discussed adequacy of employing commercially produced proficiency tests for making placement decisions. The Second Level English Proficiency Test® (SLEP), composed of reading ($k = 75$) and listening ($k = 75$) items, was administered twice as a pretest and a posttest to 487 Japanese university students. The item discrimination values showed that less than half the items did not discriminate between high and low scoring students. The result of a criterion-referenced item analysis, known as difference index (DI), indicated that students learned only one-third of the items in the program. They concluded that the SLEP® should not be used for making placement decisions because the reliability was only .81 and it had a large value of standard error of measurement. They also mentioned that the SLEP® was invalid because the test did not have a speaking section. A major goal of the program was to foster students' speaking proficiency but the test did not directly measure their speaking proficiency. They suggested that only items with a certain item discrimination value be used for making placement decisions. They also recommended the use of item response theory to make more precise placement decisions (Gorsuch & Culligan 2000).

Westrick (2005) discussed three reasons for the implementation of the Quick Placement Test-Pen and Paper Test® (QPT-PPT®) within a curriculum. The first reason was the status. The QPT-PPT® was developed by University Cambridge Local Examinations Syndicate and published by Oxford University Press, so some believed that actually using the test could somehow improve the image of their program. The second reason was that it was extremely difficult to develop in-house placement tests because of a lack of resources. The third reason was paucity of time. Usually, placement tests are administered at the beginning of an academic year, and it is difficult to find time to administer the tests in the busiest time of the year. In addition, administrators have to report the results in a short period of time in order to announce the classes in which students were placed. Westrick then reported the results of the QPT-PPT® when administered to make placement decisions. The KR20 internal consistency reliability coefficient was .66 with 120 items when 161 students took the cloze and multiple-choice tests that had reading, grammar, and vocabulary sections. He concluded that the QPT-PPT® might be effective with other groups but not for his participants and urged on the development of in-house placement tests that were connected to curricular goals and objectives.

One validation study done on an in-house placement test was by Kumazawa (2013), who evaluated the validity of an in-house placement test's score interpretations and uses based on Kane's validity framework. An interpretive argument presented for scoring, generalization, extrapolation, and decision inferences were: (a) to what extent examinees get placement items correct and high-scoring examinees get more placement items correct; (b) to what extent placement items are consistently sampled from a domain and sufficient in number so as to reduce the measurement error; (c) to what extent the difficulty of placement items matches the objectives of a reading course; and (d) to what extent placement decisions made to place examinees in their proper level

of the course have an impact on washback in the course. The backing for the score inference was based on adequate results from the item analysis. The generalization inference was supported by the composite generalizability coefficient of .92. The extrapolation inference was supported by the result of a FACETS analysis, showing that difficulty estimates of learning levels were in an expected order and course objectives for three levels were properly set in difficulty. A validity argument of decision inference was evidenced by basic-level students' score gain on an achievement test and their positive class evaluations. The validity argument presented supported the validity of the placement test score interpretations and uses.

Shizuka and Mochizuki (2014a) pointed out advantages and disadvantages of in-house and commercial placement tests. In-house placements are superior in terms of curricular coverage, but the problems are finding experts who can scrutinize the validity and students' achievement with the use of alternative test forms. An advantage of adopting commercial tests is that administrators do not have to develop tests and do the scoring. However, they could be costly and take time to administer the tests. Another problem is finding a test suitable for students' proficiency levels. In order to overcome the problems, the VELC Research Group developed several equated test forms of the VELC Test® particularly designed for Japanese university students that could be administered in 70 min at a low cost. They reported the test score reliability, criterion-related validity, and construct validity. Rasch item reliability coefficients were high, indicating item difficulty estimates were scattered along a continuum. A regression analysis showed that the test was able to account for 68 % of TOEIC® test score total variance. Compared with several competing models such as one-, two-factor confirmatory factory analysis models, the three-factor confirmatory factor analysis model best fit the data, indicating that the test gauged vocabulary, listening, and reading.

Research questions

A number of placement tests, either in-house or commercial, have been widely administered in language programs, and studies on placement tests have been done in international and Japanese settings. According to the literature (e.g., Jamieson et al. 2013; Shizuka & Mochizuki 2014a), three issues that test-users have to consider in adopting placement tests are curricular coverage, statistical property, and practicality. In-house test contents can be made to align them with program objectives, yet developing and scoring tests, and investigating test validity require a significant contribution of faculty time and expertise. If commercial tests are not as costly and if more valid test score interpretations are presented, more programs will adopt these tests and be able to use these test scores without the misuse. Thus, the purpose of this study is to evaluate VELC Test® score interpretations and uses using Kane's (2006) argument-based validity framework. To this end, the research questions for this study are:

1. To what extent do the VELC Test® items work on test-takers to make placement decisions? (Scoring)
2. To what extent is the test score generalizable to other observations so as to reduce the measurement error? (Generalizability)
3. To what extent can the test score be an indicator of what test-takers can do with their English proficiency? (Extrapolation)

4. To what extent is the test score appropriate for making placement decisions and useful for test-takers' further learning? (Decision)

Methods

Participants

In 2013, 14,245 Japanese university students in 32 different universities took the VELC Test®. Of those, 4407 students who took the same test form for placement purposes in seven different universities at the end of March or beginning of April were selected as participants in this study (n = 112, 116, 211, 297, 1153, 1235, 1283). Most of them were 18 to 22 years old. Based on a regression study conducted to estimate TOEIC® test scores from the VELC Test® scores (Shizuka & Mochizuki 2014b), their estimated TOEIC® test scores based on their could be VELC Test® scores ranged from 250 to 970 (M = 530, SD = 100), showing that the sample was representative of Japanese university students.

Materials

The VELC Test® was originally designed by the second and third authors, who specialized in language testing and vocabulary acquisition; the original items were developed based on materials written by native-speakers of English who belonged to the VELC Research Group (Shizuka & Mochizuki 2014b). With the cooperation of the VELC research group members, the test was piloted on over 5000 Japanese university students and finalized for the purpose of making Japanese university students' achievement, placement and diagnostic decisions.

The test can be completed in 70 min and costs ¥800 per person. The test contains two parts, listening and reading, each composed of three sections. The test has several test forms that are equated with common items with the Rasch model. The test scores for the two parts and total range from roughly 250 to 900 in 5-point interval. All the sections have 20 multiple-choice items (k = 120).

Vocabulary size in listening is assessed in section 1, in which test-takers listen to a Japanese equivalent and four content words in English and choose a choice that has the same meaning. For example, test-takers hear a Japanese equivalent: *manabu* (学ぶ) and four choices: (a) carry (b) spend (c) learn and (d) sell. Then, they mark (c) in the answering sheet to get it correct. The vocabulary size covered in this section ranges from JACET 1000 to JACET 7000 on the JACET corpus list of 8000 basic words (JACET Basic Word Revision Committee 2003). This section is intended to measure the extent to which test-takers can identify sounds and meanings of vocabulary.

Word identifying listening ability is measured in section 2, where test-takers look at a sentence with several blanks and hear the entire sentence and choose a choice that comes into a certain blank in the sentence. For example, on the question sheet, they read: In fact, ( ) ( ) ( ) (*) ( ) ( ) ( ) ( ) and four choices: (a) all (b) though (c) must and (d) almost. Then, they hear the following sentence twice: in fact, my sister is almost as tall as me. The word that comes in the marked blank is (d) almost. This section requires test-takers to listen to natural sounds and identify a word in a sentence.

Listening comprehension is measured in section 3, where test-takers read four choices on the question sheet and listen to sentences with the last word in the sentences deleted. They choose a choice that best suits in the last part of the sentence.

For instance, they read: (a) school (b) university (c) books and (d) jobs, and listen to: In Japan today, more than 50 % of high school students go on to. They select (b) university to get it correct. Test-takers need to listen to sentences and comprehend the meaning in order to guess what comes in the last part of the sentences.

Vocabulary size in reading is assessed in section 4, where they read a Japanese equivalent and four English words and choose the word that has the same meaning. This section is based on the Mochizuki vocabulary size test and some attempts have been made to argue for validity of the test score interpretations (e.g., Koizumi & Mochizuki 2011). The vocabulary size estimated in this section is from JACET 1000 to 7000. This section estimates test-takers' receptive vocabulary size.

Sentence structure awareness is gauged in section 5, where test-takers have to read a sentence and analyze grammatical features. Test-takers read an incomplete sentence that contains four choices somewhere in the sentence and an assigned word can be inserted in one of the four choices in the sentence. They are to choose a choice that the word can be filled in to make a grammatically correct sentence. For instance, they read: today, people (a) can use the Internet (b) find it easy to (c) communicate with (d) each other, and an assigned word, who. The word, who can be inserted in (a) to make it grammatically correct.

Reading comprehension is measured in section 6, where they have to read sentences with one blank and comprehend overall meaning in order to choose the choice that best suits in the blank. For example, they read: In Japan, high school students often ____, but university students usually do not and four choices: (a) eat lunch (b) wear uniform (c) work part-time (d) study English. They have to choose (b) to get it correct.

A main feature of the VELC Test® is the e-portfolio that reports test-takers' score, learning advice, TOEIC® equivalent test score, can-do levels, and skill mastery levels. Test-takers' scores are reported online within a few days after test administration through e-portfolio (Additional file 1: Appendix A). The TOEIC® test is widely used in Japan, so a rough estimate of the equivalent test scores are also provided. Based on the test score, learning advice is provided (Additional file 1: Appendix B). If test-takers' score in part 6 is low, it suggests that they needs to practice rapid reading so that they can read fast and grasp the main idea of a passage. It also shows test-takers' can-do levels or what they can do with their English (Additional file 1: Appendix C). For example, a test-taker can listen and understand the content of materials at junior high level with 100 % accuracy. It also illustrates test-takers' mastery of skill levels (Additional file 1: Appendix D). For instance, test-takers are able to answer items related to high school level vocabulary with 87 % accuracy. If test-takers do not have a vocabulary size up to the university level, it provides learning advice such as reading newspapers in English and looking up meanings of unknown words to help increase vocabulary size.

Procedure

Prior to the test administration, test materials including audio CDs, test booklets, scoring sheets, proctor manuals were sent to the universities. Each university administered the test on a day convenient for them at the end of March or the beginning of April.

The scoring sheets were collected and sent to the VELC Test® Administration Office, where they immediately scored the sheets and uploaded the test-takers' scores on the e-portfolio.

Analysis

To compute an official VELC Test® score, the raw score is converted to the VELC Test® score. However, in this study, all the following analyses were done based on the raw data. In order to capture the overall tendency of the data set, descriptive statistics were first calculated. Item analyses including item facility (IF), item discrimination (ID), and Rasch estimates were computed to investigate how the items were working on the test-takers. Adequate IF and ID values for norm-referenced purposes should be within .30 ~ .70 and .20 or higher. Rasch infit mean-square statistics should fall within .80 ~ 1.20 (Bond & Fox 2007, p. 243). Multivariate generalizability theory (G theory) was applied: (a) to examine the amounts of variance that could be extracted for the persons (p), items (i), sections (s), and interaction facets with the generalizability study (G study) design of p• X (s°: i°); and (b) to estimate the generalizability with the decision study (D study) design of p• X (S°:I°). Generally, G theory had been used to estimate reliability of performance-based assessment, but in some studies it was also applied to estimate reliability and dependability of paper-based multiple-choice test items (e.g., Brown 1999; Zhang 2006). A pilot study was conducted on a self-report can-do questionnaire to estimate the can-do item difficulty, and the estimates were denoted on a Rasch map, showing what test-takers with a certain person ability estimate could do with their English. In the pilot phase, 550 university students took the 20-item self-report can-do questionnaire and the 120 VELC Test® items, and then the item difficulty estimates were calibrated with the Rasch model separately for the reading and listening parts. To examine how many strata items and examinees can be split, Rasch item and person separation indexes were presented for each university. Three software programs, langtest (Mizumoto & Plonsky 2015), FACETS (Linacre 2002), mGENOVA (Brennan 2001), were used for calculating descriptive statistics and classical item analysis, Rasch estimates, and generalizability coefficients.

Results

Table 1 shows the descriptive statistics for the VELC Test® used as a placement test. If scoring is done using standardized scores, excessively high or low means causing the

Table 1 Descriptive statistics for the VELC Test® used as a placement test ($N = 4407$)

Section	*k*	*M*	*SD*	Min	Max	Skewness	Kurtosis	*α*	*SEM*
1. Vocabulary listening	20	13.37	3.23	0	20	-.44	.08	.70	1.77
2. Word identifying listening	20	9.87	3.72	0	20	.26	-.71	.74	1.90
3. Listening comprehension	20	10.27	3.16	0	20	.15	-.07	.59	2.07
4. Vocabulary	20	15.19	3.49	1	20	–1.12	.87	.81	1.52
5. Grammar structure	20	14.03	3.77	0	20	-.58	-.24	.78	1.77
6. Reading comprehension	20	11.51	4.06	0	20	-.16	-.68	.78	1.90
Total	120	72.24	17.00	22	120	-.35	-.25	.93	4.50

distributions skewed is problematic when interpreting the test scores. Sections 4 and 5 related to vocabulary and grammar were relatively easy whereas sections 2 and 3 gauging listening skills were difficult for the test-takers. Distributions for sections 4 and 5 were negatively skewed due to the high means. It would be problematic when interpreting these scores using standardized scores such as T scores, but because the scores were converted to the VELC® Test scores using the Rasch model, it was not as problematic. Sectional reliability coefficients were moderate to high ranging from .59 to .81, but the coefficient for the total score was sufficient at .93 for making placement decisions. Values for the standard error of measurement (*SEM*) for the sections with lower reliability coefficients were larger, but the value for the entire test was relatively small at 4.50.

Table 2 shows the results of the classical and Rasch item analyses. Item facility values tended to be high when words were within the 1000 word level whereas the values were lower when the word levels were higher ($r = -.67$). No relationship could be seen among item facility values, Flesch-Kincaid Grade Levels, and the number of words ($r = .10, -.22$). Having a number of items with extremely high or low IF values create a skewed distribution so the items were not desirable for norm-referenced purposes. For the listening and reading parts, 23 and 30 items were outside the range of .30 ~ .70. Items with high discrimination values create spread in the test score, so the value is a crucial indicator of sound items. For the two parts, only 17 and 2 items had values below .20. Rasch separation index for the item difficulty estimates and reliability coefficient were 31.07 and 1.00, indicating the estimates were widely spread from −2.79 to 2.58 and from −2.68 to 2.32, respectively for the two parts. When item difficulty estimates are at far ends of the scale, the standard errors are large. All the estimates were precisely estimated with the small amount of error. Fit statistics for all items were within the acceptable range.

Table 3 displays the results of the multivariate G study. In the listening part, 5, 3, 17, 2, and 73% and in the reading part, 11, 3, 15, 2, and 69% of the total score variance components were due to persons, sections, items, persons-sections interaction, and persons-items-sections intersection. The largest amount of variance accounted for in both sections was the undifferentiated error. Seventeen and 15% of the item variance indicated that items differed in difficulty to a large extent. The attenuated correlation coefficient between the listening and reading scores at .94 indicated that the scores were highly correlated. The interaction effects had the largest variance, next to the item effects. Based on the multivariate D study, when the number of items was 20 per section, generalizability coefficients (G coefficient) for the two parts and composite G coefficient were .73, .86, and .89. When there were 10 and 30 each, the coefficients decreased to .62, .79, and .83 and increased to .77, .89, and .91.

Figure 2 displays person ability, item difficulty, can-do item difficulty estimates on a Rasch map. Person ability estimates were normally distributed and centered around 1.00. Some high performers with estimates over 4 logits were on top of the map. Although items below −2 logits were relatively easy for the test-takers, most of the items were contributing to estimate the test-takers' proficiency. Though the can-do item difficulty estimates on the map were not calibrated together with the items used in this study, the estimates could show what the test-takers with certain ability could do with their English. The easiest and hardest items had the estimates of −1.97 (when

Table 2 The results of the classical and Rasch item analyses ($N = 4407$)

	Listening							Reading					
Item	Frequency/ readability	IF	ID	Diff	*SE*	Infit	item	Frequency/ readability	IF	ID	Diff	*SE*	Infit
1	1	**.96**	.26	−2.82	0.08	0.9	61	1	**.95**	.30	−2.61	0.07	0.9
2	1	**.90**	.31	−1.74	0.05	1.0	62	2	**.96**	.31	−2.77	0.08	0.9
3	1	**.90**	.46	−1.78	0.05	0.9	63	2	**.85**	.48	−1.24	0.04	0.9
4	2	**.82**	.47	−0.98	0.04	0.9	64	3	**.84**	.33	−1.19	0.04	1.0
5	3	**.92**	**.11**	−1.97	0.06	1.1	65	3	.57	.37	0.41	0.03	1.0
6	3	**.83**	.45	−1.06	0.04	0.9	66	3	**.75**	.51	−0.56	0.04	0.9
7	3	.67	.27	−0.10	0.03	1.1	67	3	**.84**	.54	−1.15	0.04	0.8
8	5	.63	.32	0.10	0.03	1.0	68	4	**.81**	.41	−0.92	0.04	0.9
9	6	.53	.27	0.58	0.03	1.1	69	6	.67	.38	−0.09	0.03	1.0
10	7	.40	.32	1.19	0.03	1.0	70	8	**.19**	**.14**	2.33	0.04	1.1
11	1	**.80**	.40	−0.89	0.04	0.9	71	1	**.93**	.46	−2.17	0.06	0.8
12	2	.68	.42	−0.15	0.03	0.9	72	2	**.93**	.39	−2.15	0.06	0.9
13	2	.41	.32	1.11	0.03	1.0	73	2	**.92**	.48	−2.08	0.06	0.8
14	3	**.72**	.42	−0.37	0.04	0.9	74	2	**.92**	.47	−2.02	0.06	0.8
15	4	.60	.39	0.27	0.03	1.0	75	3	**.84**	.45	−1.14	0.04	0.9
16	4	.33	**.17**	1.50	0.03	1.1	76	4	**.92**	.40	−2.00	0.06	0.9
17	4	.61	.37	0.21	0.03	1.0	77	4	**.76**	.51	−0.64	0.04	0.9
18	5	.38	.25	1.29	0.03	1.1	78	5	.45	.37	0.96	0.03	1.0
19	6	**.96**	**.15**	−2.76	0.08	1.0	79	6	.37	**.18**	1.33	0.03	1.1
20	7	.33	**.11**	1.51	0.03	1.2	80	7	**.72**	.38	−0.36	0.04	1.0
21	6.7(13)	.46	.44	0.89	0.03	0.9	81	6.7(12)	**.85**	.36	−1.24	0.04	0.9
22	3.0(13)	**.81**	**.09**	−0.93	0.04	1.2	82	7.6(13)	**.81**	.45	−0.95	0.04	0.9
23	11.2(13)	.57	.32	0.39	0.03	1.0	83	4.6(16)	.51	.39	0.65	0.03	1.0
24	10.7(15)	.51	.34	0.66	0.03	1.0	84	8.3(16)	**.96**	.31	−2.67	0.07	0.9
25	4.7(11)	**.14**	.23	2.73	0.05	1.0	85	9.2(19)	**.89**	.20	−1.65	0.05	1.0
26	11.2(11)	.42	.32	1.07	0.03	1.0	86	6.7(19)	**.77**	.42	−0.66	0.04	0.9
27	5.8(12)	.61	**.05**	0.18	0.03	1.3	87	12.0(25)	.32	.35	1.57	0.03	1.0
28	3.8(12)	.51	**.07**	0.65	0.03	1.2	88	8.7(34)	**.80**	.35	−0.89	0.04	1.0
29	6.7(12)	**.27**	.29	1.85	0.04	1.0	89	12.0(39)	.59	.39	0.30	0.03	1.0
30	1.2(12)	**.14**	**.19**	2.73	0.04	1.1	90	8.0(11)	**.79**	.50	−0.77	0.04	0.9
31	5.8(11)	.53	.40	0.55	0.03	1.0	91	8.3(15)	.59	.45	0.31	0.03	0.9
32	9.7(12)	**.28**	.40	1.80	0.04	0.9	92	12.0(17)	**.85**	.30	−1.25	0.04	1.0
33	5.8(13)	.68	.45	−0.17	0.03	0.9	93	12.0(32)	.63	.30	0.11	0.03	1.0
34	6.7(14)	.64	.30	0.03	0.03	1.0	94	9.9(15)	**.84**	.44	−1.18	0.04	0.9
35	12.0(12)	.35	.38	1.41	0.03	1.0	95	4.9(13)	.65	.37	0.02	0.03	1.0
36	10.0(14)	.56	.37	0.41	0.03	1.0	96	12.0(10)	**.77**	.35	−0.66	0.04	1.0
37	3.8(12)	.31	.36	1.61	0.03	1.0	97	12.0(25)	.70	.41	−0.24	0.03	0.9
38	6.9(11)	**.71**	**.14**	−0.33	0.04	1.2	98	8.0(22)	.64	.36	0.05	0.03	1.0
39	7.7(12)	**.94**	**.10**	−2.37	0.07	1.1	99	12.0(30)	.41	.36	1.13	0.03	1.0
40	3.8(12)	.39	.36	1.23	0.03	1.0	100	12.0(38)	.68	.40	−0.13	0.03	1.0
41	4.0(13)	**.83**	.39	−1.07	0.04	0.9	101	7.5(28)	.65	.38	0.01	0.03	1.0
42	8.3(10)	**.85**	.30	−1.29	0.04	1.0	102	9.5(32)	**.73**	.49	−0.41	0.04	0.9

Table 2 The results of the classical and Rasch item analyses ($N = 4407$) *(Continued)*

43	10.1(36)	.55	.33	0.50	0.03	1.0	103	7.8(35)	**.89**	.41	−1.69	0.05	0.9
44	6.7(28)	.52	.31	0.60	0.03	1.0	104	12.0(45)	.64	.49	0.06	0.03	0.9
45	7.9(28)	.53	.36	0.59	0.03	1.0	105	7.0(42)	.65	.51	−0.02	0.03	0.9
46	8.1(49)	.43	.27	1.06	0.03	1.1	106	8.6(53)	.38	.28	1.25	0.03	1.0
47	12.0(42)	.44	.28	0.99	0.03	1.1	107	10.7(55)	.61	.35	0.18	0.03	1.0
48	12.0(50)	.48	**.17**	0.80	0.03	1.2	108	12.0(76)	.43	.39	1.05	0.03	1.0
49	10.7(7)	.67	.24	−0.12	0.03	1.1	109	10.6(76)	.62	.34	0.17	0.03	1.0
50	4.8(12)	.66	.31	−0.03	0.03	1.0	110	7.3(82)	.53	.33	0.57	0.03	1.0
51	5.2(15)	.56	.46	0.43	0.03	0.9	111	11.4(79)	.35	.24	1.41	0.03	1.1
52	7.0(21)	.48	**.12**	0.80	0.03	1.2	112	9.4(35)	.52	.35	0.63	0.03	1.0
53	11.0(18)	**.72**	.20	−0.35	0.04	1.1	113	5.1(36)	.70	.42	−0.27	0.04	0.9
54	5.3(21)	**.26**	**.15**	1.88	0.04	1.1	114	12.0(48)	.73	.38	−0.43	0.04	1.0
55	11.0(37)	.59	**.17**	0.31	0.03	1.2	115	10.0(52)	.40	.39	1.16	0.03	1.0
56	8.1(51)	**.22**	.20	2.17	0.04	1.1	116	7.6(58)	**.86**	.49	−1.32	0.05	0.8
57	7.0(38)	**.28**	**.14**	1.80	0.04	1.1	117	12.0(56)	.63	.43	0.10	0.03	0.9
58	6.5(46)	**.25**	**.11**	1.99	0.04	1.1	118	11.7(56)	.57	.47	0.39	0.03	0.9
59	7.4(66)	.40	**.09**	1.19	0.03	1.2	119	8.7(61)	**.27**	.24	1.82	0.04	1.0
60	7.5(55)	.57	.32	0.37	0.03	1.0	120	9.5(64)	.34	.22	1.45	0.03	1.1

Notes. Frequency/readability = for sections 1 and 4, word frequency level based on JACET8000 or for sections 2, 3, 5, and 6, Flesch-Kincaid Grade Level (the number of words), *IF* item facility (values outside of the acceptable range are in bold), *ID* item discrimination (values outside of the acceptable range are in bold), *Diff* item difficulty estimate, *SE* standard error, *Infit* infit mean square

reading junior high level materials), and 2.94 (when listening to native speakers talking to each other). Test-takers with their person ability estimate of 0.00 can understand English when reading junior high-level materials with 88 % probability and can understand English when native speakers talking to each other with 5 % probability.

Table 4 depicts Rasch item and person reliability coefficients and separation indexes for seven universities. University A had the highest average whereas university B had the lowest. Generally, standard deviation values for universities with a larger sample were higher, indicating that a variety of test-takers in terms of their proficiency levels took the test. As Rasch item and person reliability coefficients showed, the estimates

Table 3 The result of the multivariate generalizability study with the design of p• X (s° : i°)

Facet	Listening VC	VC%	Reading VC	VC%
Persons (p)	0.01333	5 %	.94537	
	0.01694		0.02410	11 %
Sections (s)	0.00709	3 %		
			0.00714	3 %
Items (i) : s	0.04121	17 %		
			0.03367	15 %
p X s	0.00616	2 %		
			0.00495	2 %
p X i : s	0.18184	73 %		
			0.15202	69 %

Note. VC variance component

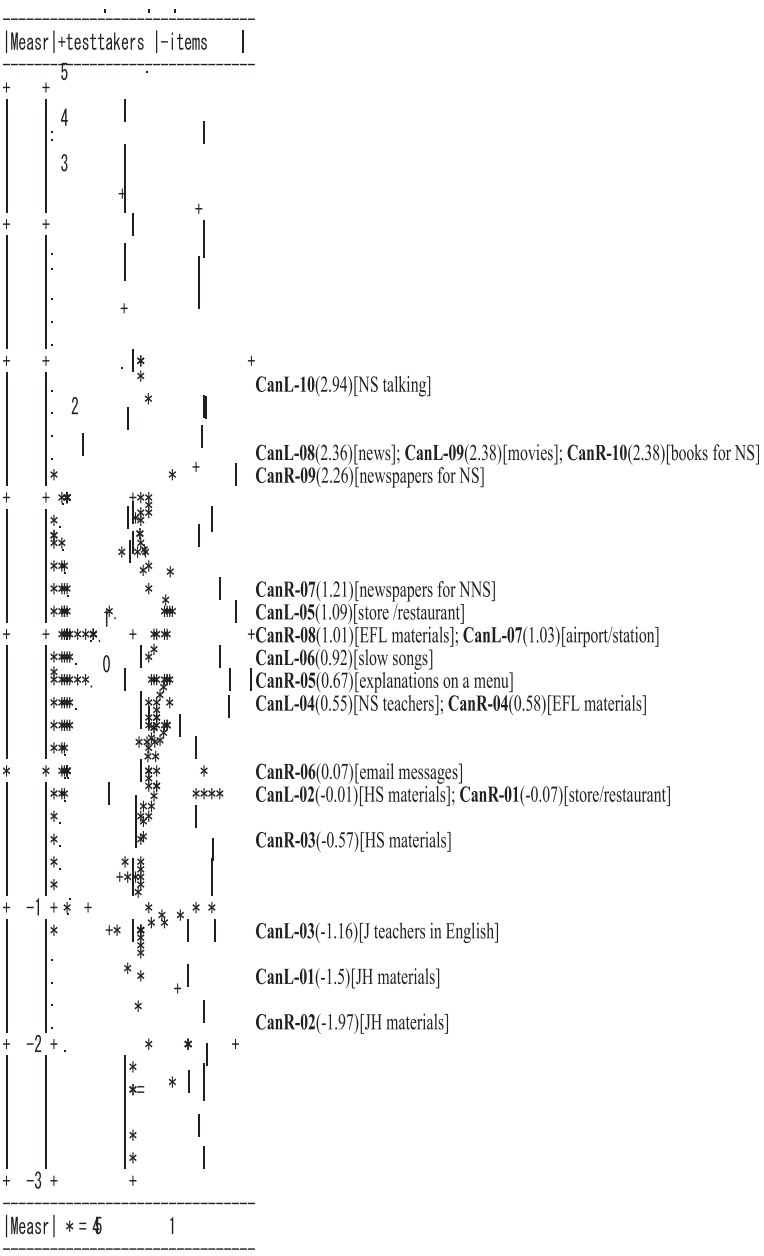

Fig. 2 Person ability, item difficulty, can-do item difficulty estimates on a Rasch map. CanL-01 = when listening to junior-high level materials; CanL-02 = when listening to high school level materials; CanL-03 = when listening to Japanese teachers giving instructions in English; CanL-04 = when taking classes taught by native speakers; CanL-05 = when salesclerks respond to your questions in a restaurant or store in a foreign country; CanL-06 = when listening to slow songs in English; CanL-07 = when listening to announcements at an airport and a station in a foreign country; CanL-08 = when watching foreign news such as CNN; CanL-09 = when watching movies in English; CanL-10 = when listening to native speakers talking to each other; CanR-01 = when reading instructions at a station in a foreign country; CanR-02 = when reading junior high level materials; CanR-03 = when reading high school level materials; CanR-04 = when reading materials written for L2 learners such as a learner dictionary; CanR-05 = when reading explanations on a menu in a foreign country; CanR-06 = when reading e-mail messages sent to you; CanR-07 = when reading news written in English for domestic readers; CanR-08 = when reading modified materials for leaners; CanR-09 = when reading news written in English for native readers; CanR-10 = when reading best-selling novels written for native readers

Table 4 Rasch Item and person reliability and separation indexes for seven universities ($N = 4407$)

University	*n*	*M*	*SD*	Item Reliability	Item Separation	Person Reliability	Person Separation
A	297	85.42	12.26	.98	6.42	.88	2.69
B	211	46.38	13.07	.96	5.24	.87	2.64
C	228	52.94	14.35	.98	6.91	.89	2.78
D	1011	77.93	15.62	1.00	14.48	.91	3.25
E	272	66.70	11.42	.98	7.77	.87	2.55
F	1235	73.34	13.40	1.00	16.86	.88	2.76
G	1153	80.18	15.25	1.00	15.07	.92	3.36

spread out along the logit scales with high reproducibility. Moderate to high Rasch item separation indexes implied that the samples were large enough to create 5 to 16 item difficulty strata. The indexes were much higher in universities D, F, and G than the others, although the standard deviation values were not as distinct. This is because the index is sensitive to a sample size. In other words, if a sample is large, the index tends to increase. Rasch person separation indexes here indicate in what levels the VELC Test® items can spread out test-takers' reading and listening abilities or, in other words, the number of distinct levels of test-takers' reading and listening abilities. The indexes were in a moderate range, indicating that the test could distinguish test-takers' proficiency into three levels.

Discussion

The research questions posed in this study were discussed based on backings collected and past findings (see Table 5). Research question 1 was: to what extent the VELC Test® items work on test-takers to make placement decisions. Placement items should measure a fairly wide range of test-takers' proficiency and discriminate their proficiency. The Rasch item separation index for the VELC Test® was 31.07, indicating that when given to the sample, item difficulty estimates spread out along the logits scale and could be split into 31 strata. Only 19 items (16 %), which had item discrimination below .20 did not discriminate their proficiency. Shizuka and Mochizuki (2014a) reported that the Rasch item separation index was 18.13 when the VELC Test® was given to 1800 Japanese university students. These results revealed that high scoring test-takers were able to select the correct answer choice, and the VELC Test® items were targeted for a wide range of proficiency levels. Because most of the items were working for placement purposes, inference from the observation to the observed score was reasonable.

Compared to the past findings on placement tests in the Japanese context, the quality of the VELC Test® items was high. Culligan and Gorsuch (1999) reported the result of the test ($k = 150$) when given to 487 Japanese university students and only 66 SLEP® test items (44 %) were discriminating high and low test-takers' proficiency levels. Westrick (2005) found that out of 120 QPT-PPT® items, only 42 items (28 %) had sufficient discriminatory power. Because an in-house placement test was piloted and tailored to a particular group of Japanese university students in a university, only four items out of 90 items (4 %) were not discriminating in Kumazawa's study (2013).

In order to develop any test either commercial or in-house, test-developers should clearly have a particular sample in mind when designing it and pilot items on the sample. To finalize the VELC Test®, three pilot trials were conducted and analyzed using

Table 5 Interpretive and validity arguments for the VELC Test® score interpretations

Inference	Interpretive argument (assumptions)	Validity argument (warrants)	Backings collected
Scoring	• The VELC Test® items work on test-takers to make placement decisions.	• Because most of the items were working for placement purposes, inference from the observation to the observed score was valid.	• Item discrimination values • Item difficulty estimates
Generalization	• The test score is generalizable to other observations so as to reduce the measurement error.	• Because the test score was reliable and generalizable with a small amount of measurement error, inference from the observed score to the universe score was valid.	• Cronbach alpha coefficients • Standard error of measurement • Generalizability coefficients
Extrapolation	• The test score indicates what test-takers can do with their English proficiency.	• Because the test score indicated what test-takers could do with their English, inference from the universe score to the target score was valid.	• Person ability estimates in relation to can-do item difficulty estimates
Decision	• The test score is appropirate for making placement decisions and useful for test-takers' further learning.	• Because the test score can be used to place certain strata and the test score report can be useful for further learning, inference from the target score to the use was fairly valid, yet the test consequence need to be investigated further.	• Rasch separation index • experts' judgment

the Rasch model. In the initial stage, a pilot test was administered to adopt anchor items that could link several test forms. The next step dealt with piloting and linking several test forms so as to revise any necessary items. The last step was to collect more data on the test forms to assure the quality of the test items.

Research question 2 was: to what extent the test score is generalizable to other observations so as to reduce the measurement error. Reliability coefficients for each part were moderate to high, ranging from .59 in part 3 to .81 in part 4. Overall, the VELC Test® was highly reliable at .93. Multivariate D study also revealed that the VELC Test® was reliable at .73, .86, and 89 respectively for the listening, reading, and both parts combined. When there were 10 and 30 items per part, G coefficients for the listening, reading, and both parts were .62, .79, and .83; and .77, .89, and .91. Values for standard error of measurement were small, indicating that test-takers' score would fall appropriately within two points. Culligan and Gorsuch (1999) and Westrick (2005) reported that reliability coefficients of the SLEP® and QPT-PPT® were .81 ($k = 150$), and .66 ($k = 120$). These results indicated that the VELC Test® could reliably estimate Japanese university students' proficiency. Thus, because the VELC Test® score was reliable and generalizable with a small amount of measurement error, inference from the observed score to the universe score was valid.

Generalizability study could reveal sources of variance in the test score. In this study, variance components for the persons (p), sections (s), items (i:s), p X s interaction, and p X i:s interaction effects were 5, 3, 17, 2, and 73 %; and 11, 3, 15, 2, and 69 % for the listening and reading parts. Because the large amount of variance was caused by

undifferentiated error, a number of factors were contributing to the total score variance of the VELC Test®. The items variance showed that items varied in difficulty to some extent. Sections differed in difficulty due to 3 % of the total variance. That was also apparent in a gap between means of sections 2 and 4 ($M = 9.87$ and 15.19). The persons variance for the reading part was larger than that of the listening part, indicating that the test-takers' reading ability varied more on the VELC Test®. Two studies which G theory was applied to analyze multiple-choice items were Brown (1999) and Zhang (2006), who reported variance components of the TOEFL® ($k = 114$) and TOEIC® ($k = 200$) tests given to 15,000 and 90,312 test-takers. Both of the studies reported that the undifferentiated variance effects accounted for over 75 %. A difference in the amount of variance was in the subtests effect. No difference in difficulty of the listening and reading parts on the TOEIC® test was observed. Even compared to these studies, about the equal amounts of variance components were observed in this study.

Research question 3 was: to what extent can the test score be an indicator of test-takers' English proficiency. To begin with, target domains that the VELC Test® covered were discussed. Shizuka and Mochizuki (2014a) reported that a confirmatory factor model that had three latent variables comprising of sections 1 and 4 as vocabulary, sections 1, 2, and 3 as listening, and sections 5 and 6 as reading best fit the data. TOEIC® test scores as a criterion measure had a high correlation coefficient with the VELC Test® scores. In this study, because the infit values for all the items fell within the acceptable range, unidimensionality of the test was maintained. In the second pilot trial, the VELC Test® and can-do statements ($k = 20$) were given to 550 Japanese university students and item difficulty estimates of the can-do statements were calibrated. The result showed that items related to reading and listening to junior high school materials were easy whereas ones related to reading books written for native speakers and listening to native speakers talking to each other were difficult to do for the test-takers. The difficulty order was reasonably ranked along the scale. In this study, by comparing item difficulty estimates of the statements to test-takers' person ability estimates, the VELC Test® scores were able to show what they could do with their English. Test-takers with their person ability estimates around 1.00 could read materials written for EFL learners and understand announcements in English-spoken countries. Because the test score indicated what test-takers could do with their English, inference from the universe score to the target score was valid. However, further study should be done to examine test-takers' actual performance in relation to their self-reported can-do statement scores.

Research question 4 was: to what extent the test score is appropriate for making placement decisions and useful for test-takers' further learning. It is common at Japanese universities that a proficiency-based program is divided into three levels based on students' placement test scores. Rasch person separation indexes were computed for the seven universities in this study and the indexes ranged from 2.64 to 3.36. This implied that the VELC Test® could be used to place test-takers into three levels. Thus, the VELC Test® scores can be used to make placement decisions within an institution.

Test-developers of the VELC Test® carefully considered ways to make the test score report useful for test-takers' further learning. First, test-takers' VELC Test® scores on

the listening and reading sections, and estimated TOEIC® test score were reported (Additional file 1: Appendix A). Their scores were explained in terms of their proficiency and achievement levels for each section. Feedback on their achievement levels and ways to improve the levels was given (Additional file 1: Appendix B). Test-takers' can-do levels roughly indicated what they could do with their English (Additional file 1: Appendix C). If a test-taker was identified as a low achiever on an item related to comprehension of junior high school materials, one would realize one had to study the materials. Diagnostic feedback on test-takers' skill mastery was given to illustrate what skills the VELC Test® was designed to measure and to the extent they had mastered. By referring to their skill mastery levels, test-takers could learn what skills they had to work on. Teachers could also gather information on what their students needed to study and design course objectives accordingly. Because the test score could be used to place certain strata and the test score report could be useful for further learning, inference from the target score to the use was fairly valid. However, the test consequence should be investigated further.

Conclusion

This study evaluated an interpretive argument for the VELC Test® score interpretations and uses by providing a validity argument using Kane's (2006) argument-based validity framework. More backings were needed to make a strong warrant for the validity argument, but all of the four inferences from scoring to decision were adequately made. However, several limitations remain in this study. First, the VELC Test® has several test forms but this study has only dealt with one form. Further investigation should be done to analyze the rest of the items and robustness of the test equating. Second, test-takers' can-do levels only estimate what they can do, so observations should be done to examine their actual performance. Third, based on experts' judgments, the VELC Test® report can be useful for test-takers' further learning, but how the report has been used and how much test-takers find it useful have to be investigated in questionnaire and interview studies. Fourth, the score report shows test-takers' skill mastery levels, but the extent to which the test items are related to the skills has not been examined. In further research, cognitive diagnostic models can be used to investigate the relatedness of the items to the skills.

This study indicated that administrators could make valid Japanese university students' placement decisions with the VELC Test®. However, when interpreting and using the test scores, administrators should refer to any necessary information on the VELC Test® in order to avoid misinterpreting and misusing the test scores. In addition, administrators should occasionally administer the VELC Test® to keep track of their students' proficiency levels, to guide their students' further learning, to develop their program including course objectives, and evaluate program effectiveness.

Additional file

Additional file 1: Appendix A. Score Report on the VELC Test® e-Portfolio. **Appendix B.** The VELC Test® Feedback on Test-Takers' Proficiency. Appendix C Test-Takers' VELC Test® Can-Do Levels. Appendix D Diagnostic Feedback on Test-Takers' VELC Test® Skill Mastery. (docx 891 kb)

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
All authors read and approved the final manuscript.

Author details
[1]Kanto Gakuin University, Odawara, Kanagawa, Japan. [2]Daito Bunka Univeristy, Itabashi, Tokyo, Japan. [3]Reitaku Univeristy, Kashiwa, Chiba, Japan. [4]Kansai University, Suita, Osaka, Japan.

Received: 12 November 2015 Accepted: 28 December 2015
Published online: 26 January 2016

References

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

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

Bachman, L.F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*, 1–34. doi:10.1207/s15434311laq0201_1.

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

Beglar, D. (2010). A Rasch-based validation of the vocabulary size test. *Language Testing, 27*, 101–118. doi:10.1177/0265532209340194.

Blais, J., & Laurier, M. (1995). The dimensionality of a placement test from several analytical perspectives. *Language Testing, 12*, 72–98. doi:10.1177/026553229501200105.

Bond, T., & Fox, C. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). New York: Routledge.

Brennan, R.L. (2001). *mGENOVA (version 2.1) [Computer software]*. Iowa City: The American College Testing Program.

Brown, J.D. (1989). Improving ESL placement tests using two perspectives. *TESOL Quarterly, 23*, 65–83. doi:10.2307/3587508.

Brown, J.D. (1999). The relative importance of persons, items, subtests and languages to TOEFL® test variance. *Language Testing, 16*, 217–38. doi:10.1177/026553229901600205.

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

Brown, J.D., Hudson, T., & Clark, M. (2004). Issues in placement survey (NetWork#40). Retrieved from University of Hawai'i, National Foreign Language Resource Center: http://nflrc.hawaii.edu/NetWorks/NW41.pdf

Chapelle, C.A. (2008). The TOEFL® validity argument. In C. A. Chapelle, M. K. Enright, & J. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 319–352). London: Routledge.

Chapelle, C.A. (2012). Conceptions of validity. In G. Fulcher & F. Davidson (Eds.), *Routledge Handbook of Language Testing* (pp. 21–33). New York: Routledge, Taylor & Francis Group.

Chapelle, C.A., Enright, M.K. , & Jamieson, J. (2010). Does an argument-based approach to validity make a difference ? *Educational Measurement: Issues and Practice, 29*, 3–13.

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302. doi:10.1037/h0040957.

Culligan, B., & Gorsuch, G. (1999). Using a commercially produced proficiency test in a one-year core EFL curriculum in Japan for placement purposes. *JALT Journal, 21*, 7–25.

Fulcher, G. (1997). An English language placement test: issues in reliability and validity. *Language Testing, 14*, 113–139. doi:10.1177/026553229701400201.

Gorsuch, G., & Culligan, B. (2000). Using item response theory to refute placement decisions. *JALT Journal, 22*, 315–325.

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

JACET Basic Word Revision Committee. (2003). *JACET list of 8000 basic words*. Tokyo: Japan Association of College English Teachers.

Jamieson, J., Wang, L., & Church, J. (2013). In-house or commercial speaking tests: Evaluating strengths for EAP placement. *Journal of English for Academic Purposes, 12*, 288–298. doi:10.1016/j.jeap.2013.09.003.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Washington: American Council on Education.

Kane, M. (2012). Validating score interpretations and uses. *Language Testing, 29*, 3–17. doi:10.1177/0265532211417210.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*, 5–17. doi:10.1111/j.1745-3992.1999.tb00010.x.

Koizumi, R., & Mochizuki, M. (2011). Development and validation of the PC version of the Mochizuki Vocabulary Size Test. *JACET Journal, 53*, 35–55.

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. doi:10.1177/0265532213475782.

Kumazawa, T. (2013). Gakunai kaihatsu pureisumento tesuto tokuten kaishaku to shiyou no datousei no hyouka nitsuite [Evaluating Validity for In-House Placement Test Score Interpretations and Uses]. *JALT Journal, 35*, 73–100.

Linacre, J.M. (2002). *FACETS (Version 3.41) [Computer software]*. Chicago: MESA.

Llosa, L. (2008). Building and supporting a validity argument for a standards-based classroom assessment of English proficiency based on teacher judgments. *Educational Measurement: Issues and Practice, 27*, 32–42. doi:10.1111/j.1745-3992.2008.00126.x.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50*, 741–749. doi:10.1037/0003-066X.50.9.741.

Mizumoto, A., & Plonsky, L. (2015). R as a lingua franca: Advantages of using R for quantitative research in applied linguistics. *Applied Linguistics*. Advance online publication. doi:10.1093/applin/amv025
Papageorgiou, S., & Cho, Y. (2013). An investigation of the use of TOEFL® Junior Standard scores for ESL placement decisions in secondary education. *Language Testing, 31*, 223–239. doi:10.1177/0265532213499750.
Pardo-Ballester, C. (2010). The validity argument of a web-based Spanish listening exam: Test usefulness evaluation. *Language Assessment Quarterly, 7*, 137–159. doi:10.1080/15434301003664188.
Sasaki, M. (1991). A comparison of two methods for detecting differential item functioning in an ESL placement test. *Language Testing, 8*, 95–111. doi:10.1177/026553229100800201.
Shizuka, T., & Mochizuki, M. (2014a). Nihonjinn daigakusei no tameno hyojyunn pureisumento tesuto to datousei no kennsho [The development and validation of a standardized placement test for Japanese University students]. *JACET Journal, 58*, 121–141.
Shizuka, T., & Mochizuki, M. (2014b). VELC Test® for testing competency: Verification of reliability and validity: Retrieved from the VELC Test® website: http://www.velctest.org/contact/VelcTest-for-TestingCompetency.pdf
VELC Research Group. (2013). The Visualizing English Language Competency Test® [Measurement instrument]. Retrieved from http://www.velctest.org
Wall, D., Clapham, C., & Alderson, J. (1994). Evaluating a placement test. *Language Testing, 11*, 321–344. doi:10.1177/026553229401100305.
Westrick, P. (2005). Score reliability and placement testing. *JALT Journal, 27*, 71–94.
Zhang, S. (2006). Investigating the relative effects of persons, items, sections, and languages on TOEIC® score dependability. *Language Testing, 23*, 351–369. doi:10.1191/0265532206lt332oa.