Koizumi et al. Language Testing in Asia (2016) 6:5
DOI 10.1186/s40468-016-0027-7

Language Testing in Asia

RESEARCH Open Access

Validity evidence of Criterion® for assessing L2 writing proficiency in a Japanese university context

Rie Koizumi[1*], Yo In'nami[2], Keiko Asano[1] and Toshie Agawa[1]

* Correspondence: rkoizumi@juntendo.ac.jp
[1]Juntendo University, Chiba, Japan
Full list of author information is available at the end of the article

Abstract

Background: While numerous articles on Criterion® have been published and its validity evidence has accumulated, test users need to obtain relevant validity evidence for their local context and develop their own validity argument. This paper aims to provide validity evidence for the interpretation and use of Criterion® for assessing second language (L2) writing proficiency at a university in Japan.

Method: We focused on three perspectives: (a) differences in the difficulty of prompts in terms of Criterion® holistic scores, (b) relationships between Criterion® holistic scores and indicators of L2 proficiency, and (c) changes in Criterion® holistic and writing quality scores at three time points over 28 weeks. We used Rasch analysis (to examine (a)), Pearson product–moment correlations (to examine (b)), and multilevel modeling (to examine (c)).

Results: First, we found statistically significant but minor differences in prompt difficulty. Second, Criterion® holistic scores were found to be relatively weakly but positively correlated with indicators of L2 proficiency. Third, Criterion® holistic and writing quality scores—particularly, essay length and syntactic complexity—significantly improved, and thus are sensitive measures of the longitudinal development of L2 writing.

Conclusion: All the results can be used as backing (i.e., positive evidence) for validity when we interpret Criterion® holistic scores as reflecting L2 writing proficiency and use the scores to detect gains in L2 writing proficiency. All of these results help to accumulate validity evidence for an overall validity argument in our context.

Keywords: Validity argument, Automated essay scoring, Rasch analysis, Multilevel modeling, Holistic scoring, Essay length, Syntactic complexity

Background

Automated essay scoring systems—including Criterion®—have been extensively researched, and their applications have spread from scoring high- and low-stakes writing exams to evaluating essays in the classroom for summative and formative purposes (e.g., Elliot & Williamson, 2013; Shermis & Burstein, 2013; Xi, 2010). Although previous studies have accumulated multiple pieces of validity evidence for the interpretation and use of Criterion®, validity evidence for local users is essential to interpret and use test scores in a meaningful way. We intend to provide such evidence in the context of assessing writing proficiency at a university in Japan.

Literature review

Criterion® uses the e-rater® automated scoring system developed by the Educational Testing Service (Burstein et al., 2013). Upon submission of an essay, Criterion® instantly produces a holistic score of 1 to 6 and presents a Trait Feedback Analysis report that suggests areas of improvement in Grammar, Usage, Mechanics, Style, and Organization & Development in the form of graphs and color-coded texts. If teachers allow, a student can access to planning tools, the scoring guide, sample essays that had received scores of 2 to 6, and the Writer's Handbook in Japanese that explains difficult terms and provided good examples.

Along with the increasingly wider applications of Criterion®, numerous studies have been conducted from various perspectives, which is well summarized in Enright and Quinlan (2010). For example, they reported that machine-scored Criterion® scores were correlated highly with human scores (e.g., $r = .76$), and that machine-generated scores based on two independent essays (e.g., $r = .80$) were correlated more highly than human-generated scores based on the same essays (e.g., $r = .53$; see also Li et al., 2014).

While these types of evidence are invaluable to evaluate the validity related to Criterion® in general, test users need to evaluate its validity in their local context. Chapelle (2015) emphasized the importance of developing one's own localized validity argument considering one's test purposes and uses. For this aim, we examine the validity of the interpretation and use for assessing second language (L2) writing proficiency at a university in Japan, when the interpretation and use are made based on scores derived from Criterion®. We investigate this from three perspectives: (a) differences in Criterion® holistic scores due to prompts (prompt difficulty), (b) relationships between Criterion® holistic scores and indicators of L2 proficiency, and (c) longitudinal changes in Criterion® holistic and writing quality scores. These areas are related to three types of inferences and are crucial in our context. This is because we are interested in comparing scores derived from different prompts, interpreting scores as indicators of L2 writing proficiency, and examining score changes before and after instructions. While these three perspectives are part of many types of validity evidence, providing them would be a step forward to a convincing validity argument (see Bachman & Palmer, 2010; Chapelle et al., 2008, 2010; Kane, 2013, for comprehensive validity argument-based frameworks).

With regard to score differences due to prompts, Weigle (2011) examined effects of the Test of English as a Foreign Language (TOEFL) Internet-based test (iBT®) independent task prompts and rater types (human scoring vs. machine scoring done by the same e-rater® engine as Criterion®) on holistic scores. She found nonsignificant and negligible effects of prompts and an interaction between prompts and rater types and a significant but small effect of rater types (partial $\eta^2 = 0.003$, 0.001, and 0.030, respectively). While the task prompts she used were similar to Criterion® prompts, her participants were 386 L2 English learners at universities in the U.S. with a small number of Japanese learners of English (6.48 %, 25/386), and their overall proficiency levels seemed higher than our students. Additionally, two prompts may not be sufficient in number, and more prompts are needed for rigorous investigation. Furthermore, other L2 writing studies such as Nagahashi (2014) and Cho et al. (2013)

demonstrated that the difficulty of different prompts vary, which necessitates more investigation on this topic.

Relationships between Criterion® scores and indicators of L2 proficiency have also been examined. Weigle (2011) reported correlations of automated e-rater® scores with self-evaluation, teacher evaluation, and writing scores based on essays written in nontest contexts of 368 students at universities in the U.S. It was found that the strength of correlations were small in most cases, with the highest correlation (r = .41) between automated scores and L2 writing language-related scores based on essays written in English courses (in contrast with courses of their major). There was also a small correlation (r = .36) between automated scores and self-assessment of L2 writing ability. Enright and Quinlan (2010) reported moderate correlations between automated e-rater® scores and TOEFL iBT® scores of integrated writing, reading, listening, and speaking (r = .59 to .61; participants' details are not reported). These correlations were not high, but were similar to those between single-human ratings and indicators of L2 proficiency (r = .56 to .61).

Lastly, longitudinal changes in Criterion® holistic and writing quality scores have been investigated across time points before and after the L2 instruction by studies using Criterion® holistic scores and/or other scores (see Table 1). Ohta (2008a), Hosogoshi et al. (2012), and Tajino et al. (2011) used Criterion® holistic scores, and Ohta (2008a, 2008b), Hosogoshi et al. (2012), and Li et al. (2015) used writing quality measures derived from Criterion® Trait Feedback. All the studies mentioned above reported significant improvement in Criterion® holistic and other writing quality scores. For example, Ohta (2008a) conducted a 16-week instruction using Criterion® to 43 university students in two TOEFL preparation classes in Japan and compared two essays of different prompts. The results of t-tests showed that Criterion® holistic scores and the number of words in the essays increased among students with TOEFL Institutional Testing Program (ITP®) scores of 500 or above, but that they did not increase among students with those of below 486. Ohta (2008b) used the same data as in 2008a study and analyzed the essays written by 25 students who submitted all the assignments from the viewpoint of vocabulary, accuracy, and organization. She reported significant increase of the number of words they wrote and improvement in overall organization.

Previous studies, as can be seen in Table 1, have provided valuable insights into the capability of Criterion® in detecting changes in writing. However, two points need to be noted. First, all previous studies had only two time points to collect data. It is preferable to measure writing three or more times, which would enable us to examine clearer patterns of score change over time and obtain stronger evidence to argue for the utility of Criterion® as a sensitive measurement tool for detecting long-term changes in L2 writing proficiency.

Second, all the previous research has used repeated t-tests or analyses of variance (ANOVAs), sometimes along with effect sizes. Some previous studies (e.g., Ohta, 2008a) did not consider a nested structure of their data in which students belong to different classes. Data are nested when data at lower levels are situated within data at higher levels. For example, in longitudinal analysis, scores are nested within students, and students are usually nested within classes. In previous studies, student data from different classes were combined into one group for analysis. However, this

Table 1 Previous studies comparing Criterion®-generated scores before and after L2 instructions

Study	Context; variable	Instruction length (Nested?)	Prompts of essays compared	Use of Criterion® holistic scores	Analysis of writing quality (Use of Criterion® and/or other measures)	Statistical analysis (time points)	Results of comparison between pretest and posttest
Ohta (2008a) [43]	University in Japan; instruction	16 weeks (yes)	Different	Yes	Yes (Criterion®)	*t*-test (2)	Higher scores & longer essays for students with TOEFL ITP® scores of 500 or above
Ohta (2008b) [25]	University in Japan; instruction	16 weeks (yes)	Different	No	Yes (Criterion® + other measures)	*t*-test (2)	Longer essays; better organization
Hosogoshi et al. (2012) [74[a]]	University in Japan; peer feedback + revision	2 weeks (yes)	Same: first vs. revised essays	Yes	Yes (Criterion® + other measures)	ANOVA (2)	Higher scores; fewer mechanical errors; longer essays; more syntactic complexity, better organization
Tajino et al. (2011) [97]	University in Japan; instruction	2 weeks (unknown)	Different	Yes	No	*t*-test (2)	Higher scores
Li et al. (2015) [70]	University in the U.S.; instruction	Unknown (no)	Same: first vs. revised essays	No	Yes (Criterion®)	*t*-test (2)	Fewer errors (with some variations among prompts and teachers)
Current study (Study 2) [81]	University in Japan; instruction	28 weeks (yes)	Different	Yes	Yes (Criterion®)	Multilevel modeling (3)	

Note. Nested? = Yes, if there is a nested structure in the data in terms of the classes or grades to which the participants belonged. [ ] = *N*. [a]They selected and analyzed those whose holistic scores increased by one or more point out of six

approach ignores plausibly unique characteristics of classes—for example, that they are proficiency-streamed or have different emphases in teaching or group dynamics caused by individual differences and interactions among class members—so it is expected that students in the same class behave more similarly than those in different classes. This analytical problem can be avoided by modeling a nested structure of contexts where students are situated under a class. This can be achieved by using generalized linear mixed modeling or, more specifically, multilevel modeling, which is also called hierarchical linear modeling, mixed-effects modeling, random-effects modeling, and others (see Barkaoui, 2013, 2014 and Cunnings, 2012, for other features and advantages of this method). We use the term "multilevel modeling" hereafter.

Current study

To provide validity evidence for the interpretation and use of Criterion® as an assessment tool, we address the following three research questions.

1. To what extent are Criterion® prompts similar in terms of difficulty? (Generalization inference)
2. Are Criterion® holistic scores positively related to indicators of L2 proficiency? (Extrapolation inference)
3. Can Criterion® holistic and writing quality scores show changes in writing over time? (Utilization inference)

The three research questions were categorized according to Chapelle et al.'s (2008) and Xi (2010) argument-based validity framework, which consists of Domain Definition (or Representation), Evaluation, Generalization, Explanation, Extrapolation, and Utilization inferences, three of which were relevant to the current study. In the argument-based validity framework, researchers first formulate an interpretive argument (or a framework to examine a test and provide justification for test interpretation and use), which includes inferences, warrants, assumptions, backing, and rebuttal. To claim that a test is sufficiently valid to interpret and use the test scores for its intended purposes, we need to clarify inferences and each inference needs to be supported by a warrant or "a law, generally held principle, rule of thumb, or established procedure" (Chapelle et al., 2008, pp. 6–7). There are assumptions behind the warrant, and each assumption is supported by adequate positive evidence (i.e., backing) or questioned by negative evidence (i.e., rebuttal). After setting the interpretive argument, researchers perform logical and empirical investigations and obtain backing and/or rebuttal. They then evaluate the interpretive argument structure along with the backing and rebuttal obtained, and make a validity argument in the test users' context (see Kumazawa et al., 2016, and Koizumi et al., 2011, for an overall validation procedure). In our case, we focus on three inferences, each of which has one warrant, one assumption, and one piece of evidence to justify the inferences and to eventually make the validity argument in our context.

Research Question 1 is related to the Generalization inference in the validity framework because prompts of similar difficulty can be used as parallel prompts and lead to

the possibility of generalizing the result from one prompt to another. Although the Generalization inference is usually associated with reliability issues, it also deals with the parallel nature of tasks and test forms (Chapelle et al., 2008, p. 20).

Research Question 2 is associated with the Extrapolation inference because positive correlations support the inference of extrapolating the results of Criterion® holistic scores to L2 (writing) proficiency used in L2 use contexts.

Research Question 3 is related to the Utilization inference because if L2 writing improves over time, Criterion® scores should be able to detect the improvement and be sensitive enough to be used as a measure to show score gains of test takers. Research Question 1 is examined in Study 1, whereas Research Questions 2 and 3 are examined in Study 2.

Each of these three research objectives characterizes our study as unique, compared with previous studies. First, we use more prompts than Weigle (2011) to examine prompt differences. Second, we use several test indicators (e.g., TOEFL iBT® scores) and one nontest indicator (i.e., self-assessment) of L2 proficiency to examine correlations with Criterion® holistic scores—for example, Weigle (2011) examined only nontest indicators. Third, we examine changes in scores over three periods, using multilevel modeling to consider the nested structure of the data.

Study 1 (for Research Question 1)

Method

Participants

There were two groups of participants ($N = 363$): (a) a first- to third-year university student group ($n = 333$) who wrote on two essay prompts and (b) an external group who wrote on all four essay prompts ($n = 30$; see Table 2 for the study design). We assigned only two prompts for the university student group, to shorten the test-taking time and thus reduce the burden of the test taking. Those who did not take all assigned prompts or who were native speakers of English were not included in the analysis.

The university student group majored in medicine at a private university in Japan. They wrote Criterion® essays for the first time as part of their English lesson. After a teacher explained characteristics and procedures of Criterion®, they wrote on two prompts in one lesson.

Table 2 Study design of Study 1

	Prompt 1: Successful students	Prompt 2: Important animal	Prompt 3: Living longer	Prompt 4: Prepare for a trip
First-year students ($n = 108$)	✓	✓		
Second-year students ($n = 111$)	✓		✓	
Third-year students ($n = 114$)	✓			✓
External group (n = 30)	✓	✓	✓	✓

Note. Essay writing conditions: 30-min time limit; show warning when 5 min remain; Spell Checker available; allow students to make a plan before working on Criterion®; limit students to 1 submission (so they can read feedback but not revise the essay)

The external group did not belong to the university to which the university student group was affiliated. This group was composed of adult Japanese learners of English, including 13 undergraduate and 10 graduate students, 6 English teachers, and 1 professional who used English for business, with a wide range of proficiency from beginning to advanced levels; their initial Criterion® holistic scores ranged from 1 to 6 (the whole score range). This group was recruited to participate in the study and given a 2000-yen prepaid card upon completion of the task. This group was included to stably equate the scores on the same scale using Rasch analysis (see Kolen & Brennan, 2004, for test equating). Each external group member read the instructions for Criterion® and completed the four essays at their own pace within a week.

Instruments and procedures

The Criterion® tests were all timed (30 min), and the participants were not allowed to use dictionaries or ask for help. Prompts were selected from expository mode prompts in the TOEFL level category in the topic library provided in Criterion®. All the participants wrote the Prompt 1 essay first. In the external group, the order of Prompts 2 to 4 was counterbalanced to avoid order effects; Prompt 1 was not counterbalanced to align with the condition of the other group.

Analysis

To examine Research Question 1, we analyzed the participants' Criterion® holistic scores. We used a concurrent calibration equating method with Rasch analysis (with Facets, Version 3.71.4; Linacre 2014), which estimates students' ability and prompt difficulty at the same time on a single, comparable scale (see Bond & Fox, 2015 and Eckes, 2011, for the Rasch model).

We then examined whether there were misfitting tasks (in this case, prompts) or persons using an infit mean square between 0.5 and 1.5 as the benchmark for tasks and persons fitting the Rasch model (Linacre 2013). We did not consider it problematic if there was an overfit (i.e., an infit mean square of below 0.5), because this shows that the tasks and persons fit the model better than expected and therefore are redundant. We did not also consider problematic an infit mean square between 1.5 and 2.0, because the item or person is "unproductive for construction of measurement, but not degrading" (Linacre 2013, p. 266). We examined response patterns when the value exceeded 2.0 because such a task or person "distorts or degrades the measurement system" (p. 266).

Results

An analysis of 363 participants' data shows there were no misfitting prompts but some misfitting (all underfitting) test takers: 14 test takers fit the Rasch model, 244 test takers (67.22 %) had infit mean squares of less than 0.50, 66 test takers (18.18 %) had infit mean squares of over 1.50, and 39 test takers out of 66 (10.74 %, 39/363) had infit mean squares of over 2.00. Close inspections of unexpected responses (reported if standardized residual ≧ 2.00) from underfitting candidates ($n = 21$) indicated that responses related to Prompt 1 seemed problematic: 19 had very different scores between Prompt 1 (the first prompt) and the second

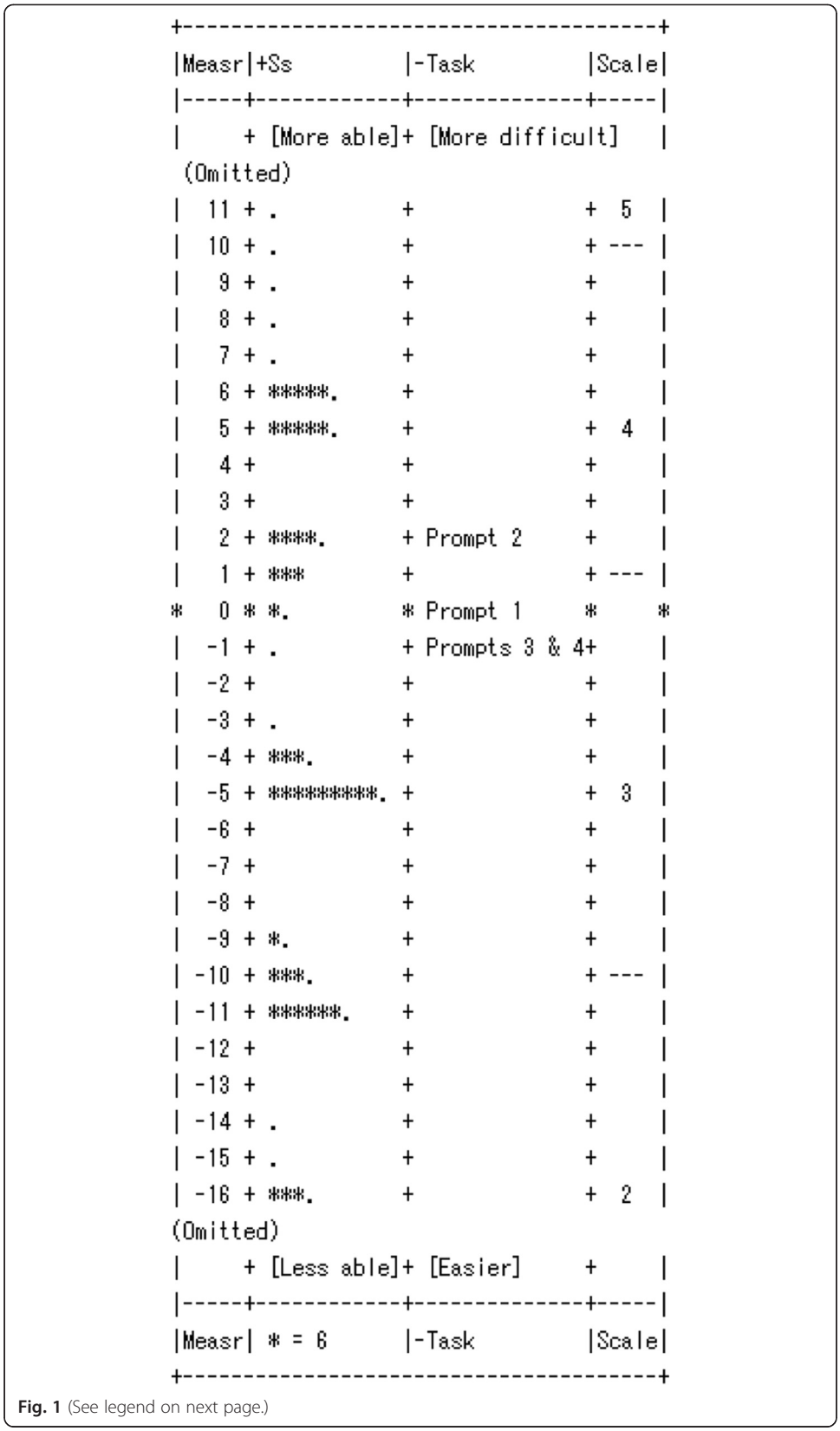

Fig. 1 (See legend on next page.)

(See figure on previous page.)
Fig. 1 Wright map for participants and prompts ($n = 348$). Ss = Participants; * = 6 participants; . = 1 participant. Seventeen participants with measures of above 11.00 and 10 participants with measures of below −16.00 were omitted from the figure. Higher values mean higher ability in the second column and more difficult prompts in the third column

prompt; 14 out of 19 had lower scores in Prompt 1 than in the second prompt and they seemed not to have used their proficiency to the fullest, due to the lack of motivation or unfamiliarity with the Criterion® procedures. Five participants out of 19 had higher scores in Prompt 1 and they seemed not to have used their proficiency to the fullest in the second prompt probably because they felt tired after writing Prompt 1. All these unexpected responses from university students ($n = 15$) were excluded (because excluding Prompt 1 leaves only another prompt, which does not much contribute to the estimation), whereas only responses to Prompt 1 were excluded from the external group ($n = 4$).

After the reanalysis ($n = 348$), there were still 256 overfitting test takers (73.56 %) and 47 underfitting test takers (13.42 %) with infit mean squares of over 1.50. We had 22 underfitting test takers (6.29 %) with infit mean squares of over 2.00 but they had infit *z*-standardized values of less than 2.00 except for one participant, which suggests no problems. Among the underfitting test takers, most (91.49 %; 43/47) were university students. One of the reasons of underfit may be because they wrote only two prompts; minor differences between two responses from the test takers seem to have been detected as misfits.

In this analysis of 348 test takers' data, 89.40 % of the score variance was explained by Rasch measures, which suggests strong unidimensionality, which is one of the assumptions in Rasch analysis (see Fig. 1 for the relationship between participants' ability and prompt difficulty on the logit scale).

We found that person reliability and prompt reliability were high (.72 and .95). According to Bond and Fox (2015) and Linacre (2013), person reliability is conceptually the same as internal consistency in classical test theory (CTT), as often reported by Cronbach's alpha. It demonstrates how varied test takers' responses are and to what extent the ordering of test takers is consistent in terms of ability. Prompt reliability has no equivalent concept in CTT and demonstrates how varied prompts are and to what extent the ordering of prompts is consistent in terms of difficulty. We can interpret that the higher both reliabilities are, the better. Table 3 shows Observed averages, that is, the

Table 3 Task measurement report in Study 1

Prompt	Total count	Observed average	Fair average	Measure (logit)	Model *SE*	Infit MnSq	Lower measure	Upper measure
2	131	3.53	3.01	1.52	0.28	0.86	0.97	2.07
1	337	3.18	3.03	0.42	0.17	0.93	0.09	0.75
3	133	3.46	3.08	−0.59	0.26	0.94	−1.10	−0.08
4	139	3.24	3.16	−1.35	0.24	1.07	−1.82	−0.88
Mean	185	3.35	3.07	0.00	0.24	0.95	–	–
SD[a]	87.8	0.15	0.06	1.08	0.04	0.07	–	–

Note. SE = Standard error. MnSq = mean squares. [a]population. Separation = 4.31; Reliability = .95. Model, Fixed (all same) chi-square = 70.8 ($df = 3$), $p < .01$

means of the raw scores of each prompt, whereas Measures show prompt difficulty values on the Rasch logit scale (with the mean of 0 and with positive values indicating more difficult prompts) after the Rasch model took participants' ability into account. Fair averages indicate prompt difficulty values converted into the original scale from 1 to 6. Due to the adjustment to the scores, Prompt 1 was the most difficult when we looked at the observed average, whereas it was the second most difficult in the fair average.

The difficulty estimates of Prompts 1 to 4 varied from −1.35 to 1.52 ($M = 0.00$, $SD = 1.08$) on the logit scale, while participants varied substantially (Measure $M = -2.46$, $SD = 9.51$; this is not shown in Table 3). Since all test takers wrote on Prompt 1 (Total count = 337), standard error was lower ($SE = 0.17$). Prompt 4 was the easiest (−1.35), followed by Prompts 3 (−0.59), 1 (0.42), and 2 (1.52) in the order of difficulty. A significant fixed chi-square value indicates that prompts were statistically different. Separation was 4.31, meaning four prompt difficulty could be separated into four groups. Using a formula "Measure ±1.96* Model *SE*," we calculated 95 % confidence intervals (CIs) of the measure (see Columns 9 and 10). For example, Prompt 2 had a CI of 0.97 to 2.07, meaning that when we make 100 trials, at 95 times, a range of 0.97 to 2.07 includes the actual Prompt 2 difficulty. Overlaps of these CIs indicate the following order of prompt difficulty, Prompt 2 > Prompt 1 > Prompts 3 and 4, with Prompt 2 being the most difficult prompt. However, when we look at the fair averages, the difference between Prompts 2 and 4 were minor (3.16 – 3.01 = 0.15). Still, a small difference may sometimes influence the results when we discuss minor differences, so this will be taken into consideration in Study 2. In sum, Study 1 shows that while the four prompts differed in difficulty, the differences were minor. This is positive evidence of validity and suggests the high generalizability of students' writing proficiency across tasks.

Significant but only small differences across prompts accord well with Weigle (2011), which we reviewed above. Furthermore, the difficulty of Prompt 1 may also have been affected by the order of prompts because all test takers wrote on Prompt 1 first, while Prompts 2 to 4 were counterbalanced in the external group, and the university student group wrote the essays on one of the three prompts as the second prompt (see *Instruments and procedures*).

Study 2 (for Research Questions 2 and 3)

Method

Participants

We analyzed data from 81 participants who were first-year university students majoring in medicine at a private university in Japan, and wrote on two essay prompts on three occasions (see Table 4). The test data in Time 1 was also used for Study 1.

Table 4 Study design in Study 2 ($N = 81$)

	Prompt 1: Successful students	Prompt 2: Important animal	Prompt 3: Living longer	Prompt 4: Prepare for a trip
Time 1: Pretest, May	✓	✓		
Time 2: Posttest 1, July	✓		✓	
Time 3: Posttest 2, December	✓			✓

Note. Prompt 1 was not analyzed in Study 2

Instructions

The 81 students took the TOEFL ITP® test in April and were placed into five proficiency-stratified courses. They took three required courses (each consisting of a 90-min class per week) of general English for 9 months (from April to January): two courses focusing on receptive and productive skills, respectively, and one course for preparing for the TOEFL ITP® and iBT®. All teachers were allowed to conduct classes according to their teaching principles.

The students took the TOEFL ITP® test twice to assess the growth of their L2 English proficiency, as well as for administrators and teachers to evaluate the effectiveness of the English program and to place the students into English proficiency-stratified classes. Additionally, the students needed to obtain a TOEFL ITP® score of 475 or above OR a TOEFL iBT® score of 53 or above to advance to the second year. Thus, they were motivated to meet the requirement and increase their proficiency. One of the goals of English language education at the university was to foster future doctors' English proficiency so that they could go abroad for clinical training and perform well in medical examinations in Japan or abroad.

The study was conducted in a naturalistic classroom environment using intact classes of TOEFL preparation courses. The students belonged to one of the five English proficiency-stratified classes (Classes A to E, with Class A aimed at the most proficient students) determined by their TOEFL ITP® test scores. They not only wrote the essays as a test, but also used Criterion® as a learning tool. This course was taught by four Japanese teachers (one for each class, with the same teacher assigned to both C and D). The students received writing instruction using Criterion® for 28 weeks in order to prepare for the TOEFL iBT® writing section. Criterion® was selected for use because (a) the same scoring engine e-rater® is used for the TOEFL iBT® writing section and Criterion®, (b) the students could practice with the same task format in Criterion® that is used in the independent writing tasks in TOEFL iBT®, and (c) Criterion® offers efficient and consistent feedback. Criterion® was used from May to December in 2013, with a 1-month summer vacation interval. The students were also encouraged to use it outside of class.

All the teachers in the Criterion® course were encouraged to use the same PowerPoint slides for the instruction prepared by the teacher of Class E. The slides began with a description of the features of Criterion®, how to write and submit essays, how to read feedback from Criterion® and their teacher, and how to revise the essays based on the feedback. While the teachers were free to set assignments by themselves, most used the assigned tasks set by the Class E teacher, who allowed a maximum of five revisions (in addition to the original submission). After the instruction, we asked the teachers via e-mail what aspect they had focused on and how they had carried out the instruction.

To characterize each class instruction, we coded the number of prompts the students wrote, the number of revisions they made, and the amount of feedback they received from their teachers. These codings were conducted for each student, using the information recorded in Criterion® and a teacher survey. Two of the authors independently coded the data of one-third of the 81 students. The agreement ratios were high, from 97.56 to 100.00 %. The remaining data were coded by the first author. Table 5 shows that the students in each class received rather

Table 5 Means and standard deviations of the number of prompts, revisions, and times of online and written feedback for each class per student in Study 2

	Class A	Class B	Class C	Class D	Class E	Total
N	17	14	18	18	14	81
TOEFL ITP® April	569.47 (21.52)	521.79 (6.89)	501.61 (6.46)	480.89 (6.11)	457.38 (9.86)	507.71 (39.73)
December	568.31 (29.88)	539.00 (18.26)	515.28 (17.98)	499.00 (23.15)	504.93 (22.90)	524.56 (34.02)
TOEFL iBT® total	70.88 (16.73)	64.43 (9.12)	56.22 (8.21)	47.89 (4.79)	47.36 (10.14)	57.33 (13.82)
Writing	17.76 (3.25)	15.93 (2.62)	15.28 (2.32)	13.67 (1.94)	12.79 (3.17)	15.12 (3.14)
Prompts[a]	5.00 (0.71)	3.43 (0.76)	3.50 (0.71)	2.28 (1.02)	4.86 (1.41)	3.77 (1.38)
Revision	6.06 (2.44)	3.14 (2.21)	3.89 (2.59)	2.89 (2.40)	6.64 (2.02)	4.47 (2.76)
Feedback[b]	1.94 (0.43)	0.00 (0.00)	1.00 (0.00)	1.06 (0.24)	4.07 (0.27)	1.57 (1.32)

Note. [a]The number of prompts written did not include the number of prompts written on the pretest and posttests. [b]We included the feedback that was provided through Criterion® and that was written on printed sheets by each teacher, but not the feedback given orally by a teacher; while the degree of spontaneous oral feedback by each teacher was unknown, the Class E teacher reportedly spent two class periods having face-to-face conferences with every student, explaining his/her strengths/weaknesses and answering his/her questions

different instruction. For example, the Class A and Class E students were assigned the similar number of prompts and revisions (5.00 and 6.06 for Class A vs. 4.86 and 6.64 for Class E), but the Class A students received less teacher feedback than the Class E students (1.94 vs. 4.07). The types of feedback depended on the teachers. For example, some focused on organization, others focused on coherence of the argument, others mentioned both major and minor linguistic errors, and yet others focused on only major ones.

Procedures

The students took Criterion® as an exam on one of the class days before the instruction (Pretest, Time 1), after 8 weeks of instruction (Posttest 1, Time 2), and after 28 weeks of instruction (Posttest 2, Time 3).

We created two additional Criterion® accounts for each first-year student for the July and December administrations so that the students could not copy their old essays. They were discouraged from searching for similar essays online and copying them. The students knew that their Criterion® score would be part of their grade.

To assess self-assessment of L2 writing proficiency, we also administered in Times 1 and 3 a questionnaire presenting descriptions of real-life tasks and asking students to what degree they thought each statement fit to their situation on a scale of 1 to 4 (1 = *The description does not fit me at all.* to 4 = *It fits me well.*). For example, a sample item says *I can coherently write an expository writing such as the one explaining task procedures if I use vocabulary and grammar that are used in a familiar situation* (B1.1 level; Tono, 2013, p. 301). We used Can-Do statements for writing developed by the CEFR-J (Common European Framework of Reference, Japan) project members. This project segmentalized six levels into 12 levels (Pre-A1, A1.1, A1.2, A1.3, A2.1, A2.2, B1.1, B1.2, B2.1, B2.2, C1, and C2) and developed descriptors to each level and skill (Tono, 2013). We used CEFR-J descriptors because they were empirically developed for Japanese learners of English. The survey consisted of 20 items that correspond to A1.1 to C2 levels. Students answered the questionnaire after completing Criterion® essays. The results of 20 items were averaged and used for analysis.

Analysis

We used Criterion® holistic scores and Trait Feedback Analysis information, which were recorded in the Criterion® system. For Criterion® scores, we analyzed the second essay written for each occasion (i.e., Prompt 2 for Time 1, Prompt 3 for Time 2, Prompt 4 for Time 3), excluding a Prompt 1 essay to avoid the impact of prompt repetition.

To obtain Criterion® writing quality scores, we computed the following measures: (a) the number of errors per 100 words (combining errors in grammar, usage, and mechanics; lower values of this measure show higher accuracy), (b) the number of words (tokens), (c) the number of words per sentence, (d) the number of transitional words and phrases per 100 words, and (e) the number of discourse elements (i.e., introductory material, thesis statement, main ideas, supporting ideas, and conclusion; with a maximum score of 5). We considered measure values as scores and also regard each as a rough indicator of (a) accuracy, (b) essay length, (c) syntactic complexity, (d) transition, and (e) organization (see Appendix A for example essays).

To examine Research Question 2, we also used TOEFL ITP® (taken in April and December) and iBT® scores (taken between September and December, but most students took the test in November or December) and self-assessment scores (in May and December). They were employed as test and nontest indicators of L2 proficiency. Reliability of the self-assessment scores in the questionnaire was high ($\alpha = .71$ for Time 1 and .96 for Time 3), and Pearson product–moment correlations were used to examine the relationships between Criterion® holistic scores and indicators of L2 proficiency.

To examine Research Question 3, we conducted multilevel modeling. The nested (i.e., multilevel or hierarchical) structure of the data was modeled using a two-level multilevel model. The Criterion® scores and time points from Times 1 to 3 were nested within students. These variables for the participants were at a lower level and constituted the Level-1 model. The students were nested within classes. The classes were at a higher level and constituted the Level-2 model (see Fig. 2). We dummy-coded Time 1 as 0 and Time 3 as 2 and Class A—the most advanced—as 0 and Class E as 4. To examine how the Criterion® scores were predicted using time points and classes, we tested three sequential models following Raudenbush and Bryk (2002) and Singer and Willet (2003). We tested the random intercepts and random slopes of the classes, under the assumption that changes were linear.

An intercept shows the initial status of students (i.e., students' writing score in Time 1). A random intercept indicates that students' writing scores in Time 1 vary across classes and are normally distributed. In the same vein, a slope shows the rate of change (i.e., the rate of change in students' writing scores between Time 1 and Time 2, between Time 1 and Time 3, and between Time 2 and Time 3). A random slope indicates that such rates of change vary across classes and are normally distributed. If intercepts

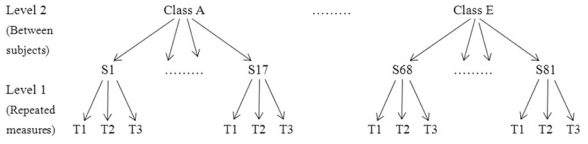

Fig. 2 Schematic explanation of the data's nested structure S = Student. T = Time (holistic or writing quality scores)

and slopes are modeled as fixed, it means that students' writing scores in Time 1 vary little across classes (everyone has a similar level of writing ability), and that the rate of change in these scores also vary little across classes (everyone improves their writing ability at the same speed). Since our participants varied in English proficiency (see Table 5) and we assumed that they were unlikely to improve their writing ability at the same speed across classes, we tested random-intercept, random-slope models. We used full the maximum likelihood estimation method available in HLM for Windows (Version 7.01; Raudenbush et al., 2011).

Results

Correlations between Criterion® holistic scores and indicators of L2 proficiency (Research Question 2)

Table 6 shows correlations with indicators of L2 proficiency assessed at the same period (see Appendix B for the whole matrix). Most of the correlations were relatively weak but positive, including the correlation between Criterion® holistic scores in Time 3 and TOEFL iBT® writing scores obtained in a similar period to Time 3 (r = .34; 95 % CI = .13, .52). This was lower than expected, but the degree of correlations were similar to Weigle (2011), which had mostly low correlations (r = .15 to .42) between automated scores and L2 nontest indicators of L2 proficiency, including self-assessment scores. The correlations were rather weak but positive, and we consider this as positive evidence. This is because regarding relationships between Criterion® holistic scores and TOEFL iBT® writing scores, one of the two tasks in the TOEFL iBT® writing was an integrated task whose features differed from Criterion® (and TOEFL iBT® independent) task. Furthermore, the actual correlations could be stronger considering errors due to a small sample size, since the upper limits of the 95 % confidence intervals were within the moderate range (i.e., .40 to .55). Therefore, Criterion® holistic scores seem to be an indicator of L2 general proficiency or writing proficiency in L2 use settings. Moreover, although our study had lower correlations than Enright and Quinlan (2010), correlational patterns were similar to theirs. Correlations between Criterion® holistic scores in Time 3 and TOEFL iBT® speaking and writing scores obtained around Time 3 (r = .34 and .38 in this study, vs. .59 to .61 in Enright & Quinlan) were a little higher than the ones between Criterion® holistic scores in Time 3 and TOEFL iBT® listening and reading scores obtained around Time 3 (r = .20 and .22 in this study, vs. .56 to .58 in Enright & Quinlan). This may indicate that Criterion® holistic scores tend to assess more productive aspects of proficiency than receptive aspects.

Table 6 Summary of correlations between Criterion® holistic scores and indicators of L2 proficiency in Study 2

	TOEFL ITP®	TOEFL iBT® R	TOEFL iBT® L	TOEFL iBT® S	TOEFL iBT® W	TOEFL iBT®	Self-assessment
Time 1 holistic scores	April .37 [.16, .54] (80)	–	–	–	–	–	Time 1 .21 [−.04, .44] (63)
Time 3 holistic scores	Dec. .32 [.11, .51] (80)	.22 [.01, .42] (81)	.20 [−.02, .40] (81)	.38 [.17, .55] (81)	.34 [.13, .52] (81)	.33 [.12, .51] (81)	Time 3 .28 [.05, .49] (70)

Note. [ ] = 95 % confidence interval. ( ) = *n*. R = Reading; L = Listening; S = Speaking; W = Writing. If r = .22 or more, $p < .05$. See Appendix B for the whole matrix. Results using Spearman's rank correlations were very similar

Changes in Criterion® holistic scores (Research Question 3)

Table 7 shows descriptive statistics of scores used for analysis. We tested three sequential models using multilevel modeling following Raudenbush and Bryk (2002) and Singer and Willet (2003). First, we tested whether the students' proficiency differed across classes—thereby testing the need to model the class variable—and whether the average student's Criterion® scores varied over time. Model 1 is called a "null model," or the "unconditional means model" in Singer and Willet's term, and is defined as follows:

Model 1:

$$\text{Level-1 Model} : \text{CRITERION}_{ij} = \beta_{0j} + r_{ij} \quad (1)$$

$$\text{Level-2 Model} : \beta_{0j} = \gamma_{00} + u_{0j} \quad (2)$$

In other words, no independent variable (i.e., time or class) was entered in Model 1. Criterion® holistic scores (CRITERION) consisted of an intercept (β_{0j}) and unmodeled variation (r_{ij}) at Level 1 (equation 1). The intercept (β_{0j}) consisted of an intercept (γ_{00}) and unmodeled variation (u_{0j}) at Level 2 (equation 2). First, the intraclass correlation in Table 8 (see the last row in Random effects) shows that 44 % of the total variance in the Criterion® scores was explained by differences among classes (0.27/(0.27 + 0.34)), exceeding the rule of thumb of 10 % and indicating the need to model the class variable. This suggests that we could not adopt Model 1, in which the variable of class was not included, because this model did not work sufficiently well, and that we needed to model class as a Level-2 variable. In other words, the remaining 56 % (100 % – 44 %) of the variance was due to Level-1 variables. Second, the intercept (γ_{00}) was 3.75 (with a standard error of 0.07) and statistically significant (see the second row in Fixed effects). This means that the average Criterion® score at Time 1 was 3.75, and this was significantly different from zero. Figure 3 shows plots of the means for each class. Although the graph appears to show some more variation across classes than the results from multilevel modeling, we can interpret the trend as mentioned above when we consider standard deviations and errors in the data.

The reliability of Model 1 was .71, indicating a substantial impact of Level 2 in Model 1; the higher the reliability, the more effect Level 2 had on Level 1. This suggests the need to model the nested structure of data (Raudenbush & Bryk, 2002). Model fit is mentioned in the explanation of Model 3 below.

Furthermore, time points were added to the Level-1 model to test whether any changes were observed in the Criterion® scores. Model 2 is called the "unconditional growth model" by Singer and Willet and defined as follows:

Model 2:

$$\text{Level-1 Model} : \text{CRITERION}_{ij} = \beta_{0j} + \beta_{1j} * \left(\text{TIME}_{ij}\right) + r_{ij} \quad (3)$$

$$\text{Level-2 Model} : \beta_{0j} = \gamma_{00} + u_{0j} \quad (4)$$

$$\beta_{1j} = \gamma_{10} + u_{1j} \quad (5)$$

The intercept (γ_{00}) was 3.46 and statistically significant, meaning that the average Criterion® score at Time 1 was adjusted to 3.46 when time was included, and this was

Table 7 Means and standard deviations across classes and time points in Study 2

Class		Holistic score	No. of errors per 100 words	No. of words	No. of words per sentence	No. of transitional words and phrases	No. of discourse elements
A		*n* = 17					
Time 1	*M*	3.76	4.51	204.35	12.47	2.63	3.94
	SD	0.66	3.40	57.98	1.86	1.14	0.66
Time 2	*M*	4.29	3.75	252.29	14.34	2.95	4.29
	SD	0.59	1.35	61.28	2.34	1.03	0.59
Time 3	*M*	4.59	2.17	284.06	17.33	2.95	4.18
	SD	0.62	1.53	60.87	5.45	1.06	0.64
B		*n* = 14					
Time 1	*M*	3.64	5.30	194.29	12.81	3.08	4.07
	SD	0.84	3.29	69.83	2.62	1.39	0.47
Time 2	*M*	3.71	5.52	213.79	13.48	3.75	3.93
	SD	0.73	2.64	74.96	1.92	2.02	0.62
Time 3	*M*	3.93	4.34	246.29	16.14	3.42	4.00
	SD	0.92	3.72	78.44	3.39	1.35	0.00
C		*n* = 18					
Time 1	*M*	3.44	6.09	209.67	13.46	3.69	4.17
	SD	0.62	2.94	53.38	2.70	1.30	0.62
Time 2	*M*	3.89	5.65	220.72	13.41	3.73	4.33
	SD	0.58	2.08	50.14	2.24	1.07	0.59
Time 3	*M*	4.06	5.56	256.67	14.33	3.80	4.22
	SD	0.54	3.35	41.98	2.79	1.13	0.55
D		*n* = 18					
Time 1	*M*	3.28	9.01	184.89	11.50	4.03	4.11
	SD	0.83	5.61	56.44	2.06	1.40	0.58
Time 2	*M*	3.61	5.30	189.56	11.18	4.20	4.56
	SD	0.61	1.69	33.15	1.54	1.18	0.51
Time 3	*M*	3.72	3.96	212.06	13.23	4.12	3.94
	SD	0.83	1.55	61.03	2.44	1.31	0.64
E		*n* = 14					
Time 1	*M*	3.00	8.89	168.64	11.23	4.26	4.21
	SD	0.88	3.59	50.74	2.22	1.87	0.58
Time 2	*M*	3.50	6.51	201.36	12.78	3.68	4.14
	SD	0.65	2.27	57.78	2.11	1.51	0.53
Time 3	*M*	3.71	5.80	223.07	12.56	4.44	4.50

Table 7 Means and standard deviations across classes and time points in Study 2 *(Continued)*

	SD	0.61	2.81	49.69	2.69	1.19	0.52
Total		*N* = 81					
Time 1	*M*	3.43	6.76	193.30	12.32	3.54	4.10
	SD	0.79	4.26	58.09	2.40	1.51	0.58
Time 2	*M*	3.81	5.30	215.88	13.01	3.66	4.27
	SD	0.67	2.16	58.78	2.28	1.40	0.59
Time 3	*M*	4.01	4.32	244.90	14.72	3.74	4.16
	SD	0.77	2.94	63.18	3.88	1.29	0.56

Table 8 Multilevel model results of holistic scores in Study 2

	Model 1 Null	Model 2 Null + time	Model 3 Null + time + class[a]
Fixed effects			
Level 1 (*n* = 243)	Coefficient (*SE*)	Coefficient (*SE*)	Coefficient (*SE*)
Intercept (γ_{00})—initial status	3.75*** (0.07)	3.46*** (0.08)	3.82*** (0.13)
Time (γ_{10})—rate of change	–	0.29*** (0.04)	0.31** (0.07)
Level 2 (*n* = 81)			
Class (γ_{01})—initial status	–	–	−0.18** (0.06)
Class (γ_{11})—rate of change	–	–	−0.01 (0.03)
Random effects			
Level 1 (*n* = 243)			
Within-student variance (r)	0.34	0.19	0.19
Level 2 (*n* = 81)			
Between-student variance (u_0)	0.27	0.38	0.32
Between-student variance (u_1)	–	0.07	0.07
Chi-square (u_0; *df*)	274.17*** (80)	280.92*** (80)	248.53*** (79)
Chi-square (u_1; *df*)	–	138.13*** (80)	137.99*** (79)
Intraclass correlation	.44		
Reliability			
Intercept (β_0)	0.71	0.71	0.67
Time (β_1)	–	0.41	0.41
Model fit			
Deviance (# of estimated parameters)	522.55 (3)	470.36 (6)	454.10 (8)
Model comparison test:			
Chi-square (*df*)	–	52.19*** (3)[c]	16.26*** (2)[d]
AIC[b]	528.55	482.36	470.18

Note. SE = Standard error. [a]Of the five classes, the highest class was coded as 0 and the lowest as 4. [b]Akaike Information Criterion (Deviance + 2*number of estimated parameters). [c]Comparison between Models 1 and 2. [d]Comparison between Models 2 and 3

The design effect for Model 1 is: 1 + intraclass correlation*([the average sample size within each cluster] − 1)) = 1 + 0.44*([243/81] − 1)) = 1.88. Values over 1 indicate the violation of the assumption of independence of observations and suggest the need to use multilevel models (e.g., McCoach & Adelson, 2010)

*$p < .05$; **$p < .01$; ***$p < .001$. These notations also apply to Table 9

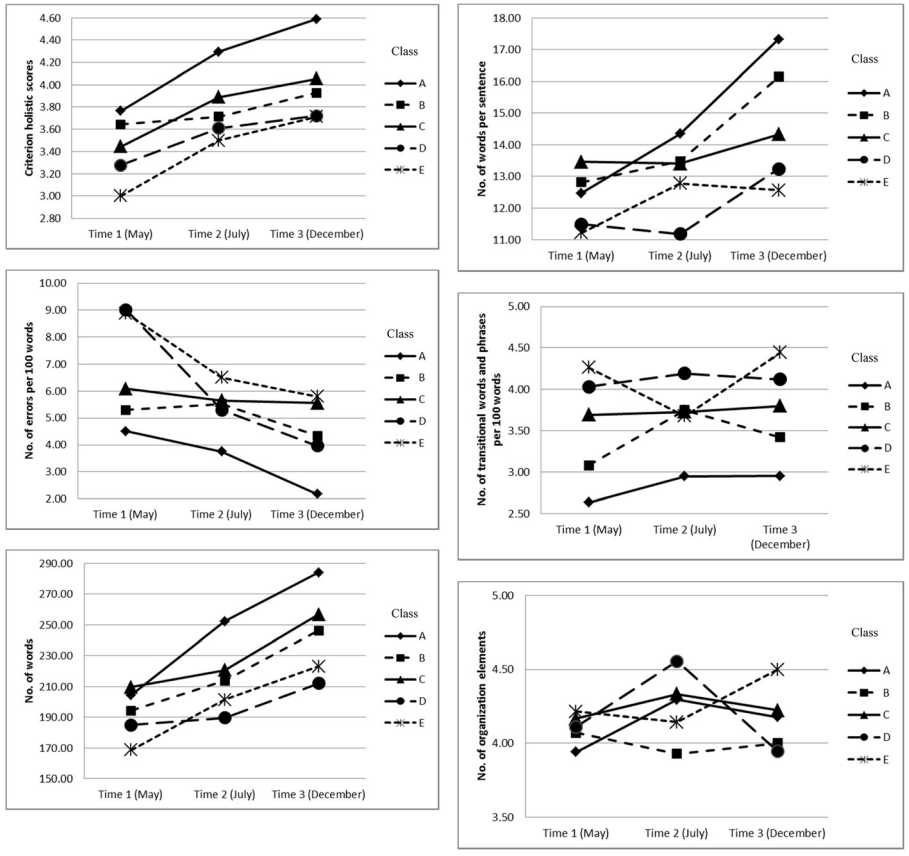

Fig. 3 Changes of the means of Criterion® holistic scores and five writing quality scores (for descriptive statistics, see Table 7)

significantly different from zero. Time (γ_{10}) was a significant predictor, and the results indicate that the Criterion® score rose by 0.29 point on average between time points. If this model is supported, this interpretation holds.

Finally, we added class to the Level-2 model to test its impact on the Criterion® scores. Model 3 is defined as follows:

Model 3:

$$\text{Level-1 Model} : \text{CRITERION}_{ij} = \beta_{0j} + \beta_{1j} * \left(\text{TIME}_{ij}\right) + r_{ij} \tag{6}$$

$$\text{Level-2 Model} : \beta_{0j} = \gamma_{00} + \gamma_{01} * \left(\text{CLASS}_{j}\right) + u_{0j} \tag{7}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} * \left(\text{CLASS}_{j}\right) + u_{1j} \tag{8}$$

A sequential comparison of the models using chi-square difference tests (e.g., $\chi^2 = 16.26$, $df = 2$, $p < .001$; see the third row in Model fit in Table 8) shows that Model 3 best explained the data. The intercept (γ_{00}) was 3.82 and statistically significant, meaning that the average Criterion® score at Time 1 was 3.82, and this was significantly different from zero. Time (γ_{10}) was a significant predictor, and the results indicate that the Criterion® score rose by 0.31 on average between time points. Besides the intercept and time, the intercept of class (γ_{01}) was a significant predictor. This indicates that the average Criterion® score for each class

at Time 1 differed by 0.18 points (e.g., if the mean of Class A—the best class—was 3.76 [see Table 7], the mean of Class B was 3.58 [3.76 – 0.18]). Further, the slope of class (γ_{11}) was not a significant predictor, indicating that the rate of change in the Criterion® score for each class did not differ and that students in all classes generally made small but similar steady progress over time, with a mean increase of 0.31 between two time points (i.e., between Times 1 and 2 and between Times 2 and 3). This means that there was a 0.62 [0.31*2] increase over the 28-week period, but initial differences in classes were retained over time.

Changes in Criterion® writing quality scores (Research Question 3)

In a similar manner to the analyses for Criterion® holistic scores, we tested sequential models for Criterion® writing quality scores. They corresponded to Models 1 to 3 above, with the only difference being that the dependent variables were not the Criterion® holistic scores, but the aforementioned five variables. Modeling each variable as a dependent variable in turn, we tested three models for each variable. Due to space limitation, we present only the model that best fit the data for each variable.

As seen in Table 9, for the number of errors per 100 words (accuracy), the best model included both time and class effects. It further indicates that the average number of errors per 100 words at Time 1 was 4.48, and that there was a difference of 1.11 between classes in Time 1. The average number of errors per 100 words did not decrease over time (nonsignificant −0.65), and the rate of such change did not differ across classes (nonsignificant −0.28).

For the number of words (essay length), the best model also included both time and class effects. This model indicates the mean at Time 1 was 208.65 words, it increased by 34.17 words over time, and there were no differences in the initial mean (nonsignificant −8.32) and change rate (nonsignificant −4.24) across classes, so students increased the number of words similarly across classes.

With respect to the number of words per sentence (syntactic complexity), the results show that the mean at Time 1 was 12.73 words, that the means increased over time by 2.08 words, and that there were no differences in the initial mean (nonsignificant −0.29). However, the change rate differed across classes (significant −0.44), indicating that Class A students increased by 2.08, Class B by 1.64 (2.08 – 0.44), Class C by 1.20 (1.64 – 0.44), Class D by 0.76 (1.20 – 0.44), and Class E by 0.32 (0.76 – 0.44).

For the number of transitional words and phrases per 100 words (transition), the best model included only the intercept, suggesting that the number of transitional words and phrases per 100 words was initially 3.65 on average, and it did not change over time.

Finally, for the number of discourse elements (organization), the best model included only the intercept, suggesting that the number of discourse elements was initially 4.18 on average and it did not change over time.

Discussion

To provide validity evidence for the interpretation and use of Criterion® for assessing L2 writing proficiency at a university in Japan, we examined three research questions. The results in relation to the validity argument are summarized in Table 10.

Table 9 Multilevel model results of the five writing quality scores in Study 2

	Error per 100 words	No. of words	No. of words per sentence	No. of transitional words per 100 words	No. of organization elements
	Null + time + class	Null + time + class	Null + time + class	Null	Null
Fixed effects					
Level 1 ($n = 243$)	Coefficient (*SE*)	Coefficient (*SE*)	Coefficient (*SE*)	Coefficient (*SE*)	Coefficient (*SE*)
Intercept (γ_{00})—initial status	4.48*** (0.56)	208.65*** (11.72)	12.73*** (0.40)	3.65*** (0.12)	4.18*** (0.03)
Time (γ_{10})—rate of change	−0.65 (0.39)	34.17*** (5.71)	2.08*** (0.43)	–	–
Level 2 ($n = 81$)					
Class (γ_{01})—initial status	1.11*** (0.26)	−8.32 (4.45)	−0.29 (0.16)	–	–
Class (γ_{11})—rate of change	−0.28 (0.17)	−4.24 (2.16)	−0.44** (0.15)	–	–
Random effects					
Level 1 ($n = 243$)					
Within-student variance (r)	4.45	962.55	3.31	1.25	0.33
Level 2 ($n = 81$)					
Between-student variance (u_0)	7.56	2361.17	2.38	0.70	0.0002
Between-student variance (u_1)	2.64	253.08	1.67	–	–
Chi-square (u_0; *df*)	246.09*** (79)	319.44*** (79)	150.75*** (79)	216.74*** (80)	71.86 (80)
Chi-square (u_1; *df*)	176.98*** (79)	123.59*** (79)	162.57*** (79)	–	–
Intraclass correlation	–	–	–	0.36	0.0006
Reliability					
Intercept (β_0)	0.67	0.75	0.46	0.63	0.00
Time (β_1)	0.54	0.35	0.50	–	–
Model fit					
Deviance (# of estimated parameters)	1193.68 (8)	2555.23 (8)	1132.04 (8)	821.30 (3)	422.00 (3)
AIC	1209.68	2571.23	1148.04	827.30	428.00

Note. **$p < .01$; ***$p < .001$. The design effect for each null model is 1.46, 2.04, 1.58, 1.72, and 1.00 (see Table 8 *Note*.) They all exceeded 1 and thus indicate the need to use multilevel models

Table 10 Inferences, warrants, assumptions, and backing in the validity argument for the Criterion®

Inference	Warrant that justifies the inference	Assumption underlying the warrant	Backing for the assumption
Generalization	Observed Criterion® scores are estimates of expected scores over the relevant parallel prompts.	Criterion® prompts are parallel in term of difficulty.	Significant but minor differences across prompts.
Extrapolation	The construct of L2 writing proficiency as assessed by Criterion® accounts for the quality of linguistic performance in L2 writing contexts.	Criterion® holistic scores are related to indicators of L2 proficiency.	Relatively weak but positive correlations between Criterion® holistic scores and indicators of L2 proficiency.
Utilization	Estimates of the writing proficiency obtained from Criterion® are useful for making decisions about long-term gains in L2 writing proficiency.	Criterion® holistic and writing quality scores can show changes in writing over time.	Significant and constant differences across three time points in Criterion® holistic scores, as well as scores of essay length, and syntactic complexity.

Research Question 1 addressed to what extent Criterion® prompts are similar in terms of difficulty. This concerned the Generalization inference in Chapelle et al.'s (2008) and Xi (2010) argument-based validity framework. We found that, despite statistically significant differences in four prompts, these differences were minor as seen in fair average scores. This can be used as backing (positive evidence) for the Generalization inference required for the validity argument and could support generalizing the students' writing proficiency across prompts.

Research Question 2 asked whether Criterion® holistic scores were positively related to indicators of L2 proficiency. This concerned the Extrapolation inference in Chapelle et al.'s (2008) and Xi (2010) argument-based validity framework. We found relatively weak but positive correlations between Criterion® holistic scores and indicators of L2 general or writing proficiency, which can be interpreted as backing for the Extrapolation inference.

Research Question 3 addressed whether Criterion® holistic and writing quality scores were able to show changes in writing over time. This concerned the Utilization inference in Chapelle et al.'s (2008) and Xi (2010) argument-based validity framework. We observed significant changes in Criterion® holistic scores and scores of text length and syntactic complexity (i.e., the number of words and the number of words per sentence), and these can be used as backing for the Utilization inference, supporting the use of Criterion® for detecting changes. Thus, scores derived from Criterion® seem sufficiently sensitive to reflect changes in students' writing proficiency in relation to L2 writing and other instruction of receptive and productive skills (see *Instructions* for details).

Additionally, the overall trend of improvement in holistic scores (with the increase of 0.62 [0.31*2 from the Time (γ_{10}) parameter in Model 3] after 28 weeks of instruction) is in line with previous studies (Hosogoshi et al., 2012; Ohta, 2008a; Tajino et al., 2011). This score change may appear small but is much larger than differences in prompt difficulty (i.e., 0.15 difference of Prompts 2 and 4) and thus indicates substantive improvement beyond measurement errors.

Regarding how writing quality scores changed in terms of accuracy, essay length, syntactic complexity, transition, and organization, the results suggest that patterns of development in writing quality vary across aspects in focus. Firstly, the number of

errors per 100 words did not significantly improve over time. The lack of improvement in accuracy did not accord well with previous studies (Hosogoshi et al., 2012; Li et al., 2015).

The essay length increased by 34.17 words on average over time across classes. The trend of writing longer essays was also reported in Hosogoshi et al. (2012) and Ohta (2008a, 2008b).

Syntactic complexity as measured by the number of words per sentence improved over time, which is in line with Hosogoshi et al. (2012) reporting a significant increase in the number of words per T-unit and that of S-nodes per T-unit. However, faster improvement in syntactic complexity for students with higher proficiency does not seem to have been reported in the literature. The fourth subgraph in Fig. 3 shows how syntactic complexity increased over time. While the Class A and B students consistently increased the complexity, the Class C and E students had a rather flat trajectory.

The number of transitional words and phrases did not increase over time. Transitional words/phrases help to improve transition in essays, and help readers understand the content easily. However, using more transitional words/phrases may not necessarily be helpful. Rather, too many transitional markers may look redundant, and are something that effective writers avoid. This may explain why no change was observed in the number of transitional words and phrases.

Organization as measured by the number of discourse elements did not improve over time, which was in contrast to Hosogoshi et al. (2012) and Ohta (2008b). The mean of 4.10 in Time 1 (see Total in Table 7) suggests that students included four elements in their essay. A close analysis of the typical essays shows that most students wrote a thesis statement, main ideas, supporting ideas, and conclusion, but introductory material was mostly missing, probably because the introductory material is difficult to write and learn. For example, the Class E teacher reported that she covered how to write introductory material multiple times by showing her students examples of good introductions, discussing the characteristics of desirable introductions, and giving written and oral feedback on writing organization. Many of her students, however, could not include the appropriate introductory material, although Fig. 3, the sixth subfigure shows nonsignificant but a small increase in the number of organization elements in Class E.

Overall, our results suggest that over the 28-week instruction, the students tended to write more words with more syntactic complexity. Along with these changes in the essay features, the students were able to attain higher Criterion® holistic scores, at least in the timed expository writing in the current study.

Conclusion

To explore the validity of the Criterion®-score-based interpretation and use for assessing L2 writing proficiency at a university in Japan, we investigated three perspectives, each related to an inference in the interpretive argument. First, we found that prompts differed in difficulty to a minor degree, as the fair average of each prompt differed little—the largest difference was 0.15 points between Prompts 2 and 4. This difference does not seem to matter in most cases, but it should be considered when interpreting small differences. Secondly, Criterion® holistic scores were found to be correlated positively with indicators of L2 proficiency, suggesting that the holistic

scores may reflect L2 proficiency in L2 use contexts. Finally, we examined whether Criterion® holistic and writing quality scores can detect changes after the 28-week instruction period. The results suggest an improvement in holistic scores, essay length, and syntactic complexity. Since we obtained backing for the three inferences, our validation of Criterion® has made substantial progress to eventually argue for the validity of the interpretation and use based on Criterion® scores in our context.

Based on these findings, pedagogical and methodological implications are offered. Firstly, pedagogically, our findings can be used as backing for supporting the validity of the interpretation and use based on Criterion® scores, as explored in the Discussion. Although this study aims to contribute to a validity argument in our local context, these findings may be able to serve as evidence or as a benchmark for comparison with other studies in similar contexts. For comparison, it should be noted that participants in Study 1 had a wide range of L2 proficiency, whereas those in Study 2 had relatively high L2 proficiency (e.g., TOEFL ITP® average in Time 1 = 507.71; see Table 5), so those in Study 2 may not be generalized to typical Japanese university students learning English as an L2.

Further, our methodological approach, namely, investigation into prompt difficulty using Rasch analysis and longitudinal examination using multilevel modeling would be helpful for other similar studies and arguably the strength of our study. According to Bond and Fox (2015), Rasch analysis enables researchers to estimate test takers' ability and task difficulty separately on the same scale and to compare the difficulty of tasks even when all participants do not take all tasks. It further provides a wide variety of rich information on test takers, tasks, and an overall test, such as person and task misfits, measurement errors for each person and task, and person and task reliability, all of which can contribute to the accumulation of validity evidence in the argument-based validity framework (Aryadoust, 2009). Multilevel modeling allows researchers to perform rigorous examinations of longitudinal data while considering the nested structure, for example, in which scores at three or more time points are nested within students, and students are nested within classes or schools. This analysis is sufficiently flexible to allow researchers to examine characteristics of data fully by building models with intercepts and slopes, both of which can be set at either fixed or random (Barkaoui, 2013, 2014; Cunnings, 2012). The use of Rasch analysis and multilevel modeling would help conduct a well-organized validation and thorough validity inquiry.

Our results need to be replicated and expanded by considering the following. First, we used a pretest-posttest design without a control group. We were not able to differentiate the effects of L2 writing instructions from the effects of L2 instructions for speaking, listening, and reading, and also students' proficiency levels, teachers' different teaching styles, and other extraneous variables. Secondly, we only analyzed holistic and writing quality scores derived from an automated scoring system. The comparison with human-rated scores would strengthen the findings and the validity argument, although previous research suggests high correlations between automated scores and human ratings (Enright & Quinlan, 2010). Thirdly, further research is needed including various essay types other than expository essays and focusing on wider aspects of writing quality using various measures. Finally, we should investigate other perspectives required to support or refute inferences and to present a convincing validity argument. Specifically, as we have provided backing for the Generalization, Extrapolation, and

Utilization inferences, we should in the future examine other areas related to the Domain representation, Evaluation, Explanation, and Extrapolation inferences, as well as testing the consequences in the Utilization inference (see Xi, 2010, for specific questions to be examined).

Appendix

Appendix A

Table 11 Typical example of writing essays that showed longitudinal improvement (written by a female student in Class A)

Time 1, Prompt 2: Important animal
The most important animal in my country,Japan,is a dog. There are two reasons.
First,dogs are the most popular animal in Japan. I think one of the animals [Possesive errors] people want to have as a pet is a dog. Many of my friends who have a pet keep a dog. In pet shops, the number of saling dogs is largest. Judging from these aspects, Japanese [Missing or Extra Articles] people are the most familiar with dogs in all animals.
Second,a dog is a very useful animal. Dogs have been kept by Japanese people since ancient times **such as** yayoi-period. No other animals is [Subject-Verb Agreement] kept by them for such a long time. In those days,dogs helped people with farming and getting foods. This fact shows that ancient people choose a dog as a useful helper. **Still**, dogs help many people in many cases. Mainly,dogs help disabled people.
For these reasons,dogs are popular and help p [Other errors]
[Organization & Development: Introductory material and Conclusion were not detected.]; [Repetition of Words: 28 words: animal(s), dog(s), people, helped]; [Passive Voice: Dogs have … yayoi-period.; No other animals … long time.]
Holistic score = 3; Errors = 3.73; Words = 134; Words/Sentence = 10.4; Transitional words = 2.99; Organization = 3

Time 2, Prompt 3: Living longer
I think the reasons that people are living longer now are the development of medicine and trend of living a healthy life.
There are many diseases that cannot be cured before World War *&#8545 [ll: garbled character]*; and can be cured completely now thanks to the development of medicine. I think most [Missing or Extra Article] famous disease which shows this is cubitous [Spelling]. **Before** the end of [Spelling]WW& *#8545 [ll: garbled character]* ; [Extra Comma], many Japanese people died of this disease. The main cause of death in Japan at that time is said to be cubitous [Spelling]. **However**, since people study hard about cubitous [Spelling], they discover the way of curing cubitous [Spelling] **and** cubitous [Spelling] became a disease which does not kill people. This discovery contribute [Subject-Verb Agreement] to saving many people's lives.
Nowadays, more and more people try to eat healthy foods and take exercise moderately. The lifestyle is very important for your health, **so** it must be related with how long people live. There are many examples which shows people consider about living a healthy life gradually. Japanese food is popular among European and American people. This is because Japanese food is said to be healthy. **In fact**, the average of life longevity [Spelling] in Japan is very high. There are many [Missing Final Punctuation]
To sum up, the development of medicine and trend of living a healthy life contribute to people's living longer.
[Organization & Development: Introductory material was not detected.]; [Repetition of Words: 10 words: people, people's]
Holistic score = 4; Errors = 5.61; Words = 214; Words/Sentence = 14.5; Transitional words = 2.34; Organization = 4

Time 3, Prompt 4: Prepare for a trip
I would take my mobile phone. Mobile [Missing or Extra Article] phone is inevitable [Missing or Extra Article] item for my life. Taking only it with me enables me to do a lot of things.
First, I can contact with my friends or my family frequently although I am away from them. **Of course**, I can see many prople [Spelling] and make new friends during my trip, but it is sure that I will feel like talking with my friend in my hometown. **In addition**, if my parents have never heard from me for as long as one year, they may be worried about my safety and health. It is essential that I send a message to my friends and family or phone them.
Second, if I always take my cell phone with me, I can take a picuture [Spelling] whenever I feel the scenes beautiful or encounter a very rare scene. Some people say that a camera can take a much better picture than a mobile pnone [Spelling]. **However**, as technology is developing, we can take by mobile phone as nice pictures as by [Preposition Error] a camera.
Third, the application included by a mobile phone shows me a map around where you are and the way to your destination. **Even if** I lose my way and cannot know which road I should take, all I have to do is to take my mobile phone from my pocket. Mobile [Missing or Extra Article] phone leads me to my destination.
For these various kinds of faculty, only [Missing or Extra Article] mobile phone helps me in many situations. Mobile phones would surely support my long trip.
[Repetition of Words: 62 words: I, take, my, mobile, phone, taking, me]
Holistic score = 4; Errors = 4.15; Words = 254; Words/Sentence = 17.1; Transitional words = 2.76; Organization = 5

Note. Underlined = parts that require some explanation. [ ] = Explanation of errors detected by Criterion®. **Bolded** = Transitional word or phrase. *Italicized* = The authors' note

Appendix B

Table 12 Correlations between Criterion® holistic scores and indicators of L2 proficiency

	M	*SD*	Time 2	Time 3	ITP® April	ITP® Dec.	iBT® R	iBT® L	iBT® S	iBT® W	iBT®	Time 1 self-assessment	Time 3 self-assessment
Time 1 holistic	3.43	0.79	.55 (81)	.47 (81)	.37 (80)	.28 (80)	.21 (81)	.27 (81)	.45 (81)	.40 (81)	.39 (81)	.21 (63)	.26 (70)
Time 2 holistic	3.81	0.67	–	.63 (81)	.40 (80)	.34 (80)	.23 (81)	.20 (81)	.48 (81)	.34 (81)	.36 (81)	.19 (63)	.27 (70)
Time 3 holistic	4.01	0.77		–	.40 (80)	.32 (80)	.22 (81)	.20 (81)	.38 (81)	.34 (81)	.33 (81)	.18 (63)	.28 (70)
ITP® April	507.71	39.73			–	.75 (79)	.50 (80)	.55 (80)	.65 (80)	.57 (80)	.68 (80)	.51 (62)	.27 (69)
ITP® Dec.	524.56	34.02				–	.42 (80)	.43 (80)	.59 (80)	.48 (80)	.58 (80)	.45 (63)	.31 (69)
iBT® R Dec.	16.15	4.59					–	.54 (81)	.46 (81)	.51 (81)	.78 (81)	.19 (63)	.26 (70)
iBT® L Dec.	13.23	5.37						–	.66 (81)	.57 (81)	.88 (81)	.19 (63)	.30 (70)
iBT® S Dec.	12.83	3.79							–	.58 (81)	.81 (81)	.35 (63)	.50 (70)
iBT® W Dec.	15.12	3.14								–	.77 (81)	.20 (63)	.16 (70)
iBT® Total Dec.	57.33	13.82									–	.29 (63)	.38 (70)
Time 1 Self-assessment	2.45	0.57										–	.43 (56)
Time 3 Self-assessment	2.75	0.56											–

Note. () = *n*. R = Reading; L = Listening; S = Speaking; W = Writing. $r = .22$ or more means $p < .05$

Acknowledgements
We would like to thank Junichi Azuma and Eberl Derek for their invaluable support for the current project. This study was funded by the Japan Society for the Promotion of Science (JSPS) KAKENHI, Grant-in-Aid for Scientific Research (C), Grant number 26370737.

Authors' contributions
RK participated in the design of the study, collected the data, performed the statistical analysis, and drafted the manuscript. YI assisted RK in performing the statistical analysis and drafted the manuscript. AK and TA collected the data and assisted RK and YI in drafting the manuscript. All authors read and approved the final manuscript.

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

Author details
[1]Juntendo University, Chiba, Japan. [2]Chuo University, Tokyo, Japan.

Received: 28 February 2016 Accepted: 22 April 2016
Published online: 04 July 2016

References

Aryadoust, SV. (2009). Mapping Rasch-based measurement onto the argument-based validity framework. *Rasch Measurement Transactions, 23*, 1192–1193. Retrieved from http://www.rasch.org/rmt/rmt231f.htm.
Bachman, L, & Palmer, A. (2010). *Language assessment in practice*. Oxford, U.K.: Oxford University Press.
Barkaoui, K. (2013). Using multilevel modeling in language assessment research: A conceptual introduction. *Language Assessment Quarterly, 10*, 241–273. doi:10.1080/15434303.2013.769546.
Barkaoui, K. (2014). Quantitative approaches for analyzing longitudinal data in second language research. *Annual Review of Applied Linguistics, 34*, 65–101. doi:10.1017/S0267190514000105.
Bond, TG, & Fox, CM. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York, NY: Routledge.

Burstein, J, Tetreault, J, & Madnani, N. (2013). The e-rater® automated essay scoring system. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 55–67). New York, NY: Routledge.

Chapelle, CA. (2015). *Building your own validity argument* (Invited lecture at the 19th Annual Conference of the Japan Language Testing Association (JLTA)). Tokyo, Japan: Chuo University.

Chapelle, CA., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language.™*. New York NY: Routledge.

Chapelle, CA, Enright, MK, & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29*(1), 3–13. doi:10.1111/j.1745-3992.2009.00165.x.

Cho, Y, Rijmen, F, & Novák, J. (2013). Investigating the effects of prompt characteristics on the comparability of TOEFL iBT™ integrated writing tasks. *Language Testing, 30*, 513–534. doi:10.1177/0265532213478796.

Cunnings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Language Research, 28*, 369–382. doi:10.1177/0267658312443651.

Eckes, T. (2011). *Frankfurt am Main*. Germany: Peter Lang. Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments.

Elliot, N, & Williamson, DM. (2013). *Assessing Writing* special issue: Assessing writing with automated scoring systems. *Assessing Writing, 18*, 1–6. doi:10.1016/j.asw.2012.11.002.

Enright, MK, & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring. *Language Testing, 27*, 317–334. doi:10.1177/0265532210363144.

Hosogoshi, K, Kanamaru, T, Takahashi, S, & Tajino, A. (2012). Eibun sanshutsu niataeru fidobakku no kouka kenshou [Effectiveness of feedback on English writing: Focus on Criterion® and feedback]. *Proceedings of the 18th annual meeting of Association for Natural Language Processing,* 1158–1161. Retrieved from http://www.anlp.jp/proceedings/annual_meeting/2012/pdf_dir/P3-22.pdf.

Kane, MT. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73. doi:10.1111/jedm.12000.

Koizumi, R, Sakai, H, Ido, T, Ota, H, Hayama, M, Sato, M, & Nemoto, A. (2011). Toward validity argument for test interpretation and use based on scores of a diagnostic grammar test for Japanese learners of English. *Japanese Journal for Research on Testing, 7*, 99–119.

Kolen, MJ, & Brennan, RL. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer.

Kumazawa, T, Shizuka, T, Mochizuki, M, & Mizumoto, A. (2016). Validity argument for the VELC Test® score interpretations and uses. *Language Testing in Asia, 6*(2), 1–18. doi:10.1186/s40468-015-0023-3. Retrieved from http://www.languagetestingasia.com/content/6/1/2/abstract.

Li, J, Link, S, & Hegelheimer, V. (2015). Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction. *Journal of Second Language Writing, 27*, 1–18. doi:10.1016/j.jslw.2014.10.004.

Li, Z, Link, S, Ma, H, Yang, H, & Hegelheimer, V. (2014). The role of automated writing evaluation holistic scores in the ESL classroom. *System, 44*, 66–78. doi:10.1016/j.system.2014.02.007.

Linacre, JM. (2013). *A user's guide to FACETS: Rasch-model computer programs (Program manual 3.71.0)*. Retrieved from http://www.winsteps.com/a/facets-manual.pdf.

Linacre, J. M. (2014). *Facets: Many-facet Rasch-measurement (Version 3.71.4) [Computer software]*. Chicago: MESA Press.

McCoach, DB, & Adelson, JL. (2010). Dealing with dependence (Part I): Understanding the effects of clustered data. *Gifted Child Quarterly, 54*, 152–155. doi:10.1177/0016986210363076.

Nagahashi, M. (2014). *A study of influential factors surrounding writing performances of Japanese EFL learners: From refinements of evaluation environment toward practical instruction. (Unpublished doctoral dissertation)*. University of Tsukuba, Japan.

Ohta, R. (2008a). Criterion: Its effect on L2 writing. In K. Bradford Watts, T. Muller, & M. Swanson (Eds.), *JALT2007 Conference Proceedings*. Tokyo: JALT. Retrieved from http://jalt-publications.org/archive/proceedings/2007/E141.pdf.

Ohta, R. (2008b). The impact of an automated evaluation system on student-writing performance. *KATE [Kantokoshinetsu Association of Teacher of English] Bulletin, 22*, 23–33. Retrieved from http://ci.nii.ac.jp/naid/110009482387.

Raudenbush, SW, & Bryk, AS. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, SW, Bryk, AS, Cheong, YF, Congdon, RT, Jr., & du Toit, M. (2011). *HLM7: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International.

Shermis, MD, & Burstein, J. (Eds.). (2013). *Handbook of automated essay evaluation: Current applications and new directions*. New York, NY: Routledge.

Singer, JD, & Willett, JB. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford, U.K.: Oxford University Press.

Tajino, A, Hosogoshi, K, Kawanishi, K, Hidaka, Y, Takahashi, S, & Kanamaru, T. (2011). Akademikku raitingu jugyou niokeru fidobakku no kenkyuu [Feedback in the academic writing classroom: Implications from classroom practices with the use of Criterion®]. *Kyoto University Researches in Higher Education, 17*, 97–108. Retrieved from http://www.highedu.kyoto-u.ac.jp/kiyou/data/kiyou17/09_tazino.pdf.

Tono, Y. (Ed.). (2013). *CAN-DO risuto sakusei katsuyou: Eigo toutatudo shihyou CEFR-J gaido bukku [The CEFR-J handbook: A resource book for using CAN-DO descriptors for English language teaching]*. Tokyo: Taishukan.

Weigle, SC. (2011). Validation of automated scores of TOEFL iBT tasks against nontest indicators of writing ability. *ETS Research Report,* RR-11-24, TOEFLiBT-15. Retrieved from http://dx.doi.org/10.1002/j.2333-8504.2011.tb02260.x

Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading [Editorial for special issue: Automated scoring and feedback systems for language assessment and learning]? *Language Testing, 27*, 291–300. doi:10.1177/0265532210364643.