Writing Assessment in University Entrance Examinations: The Case for "Indirect" Assessment

Kristy King Takagi

English language programs in Japan, and around the world, often use placement testing to gauge students' English proficiency levels and then place students into classes at those levels. English teachers and program administrators typically favor placement testing because it allows for more efficient teaching and because class placement affects and matters to students. However, placement testing for writing classes can be burdensome and time-consuming because a typical approach in language programs is to obtain writing samples from students and ask teachers to rate the samples. In light of this burden, and the problems associated with the rating of writing samples, the focus of this paper is to examine whether an objective multiple choice test of writing knowledge could serve as a supplement to or substitute for the typical rating of writing samples for English writing class placement.

As Hamp-Lyons (1991) pointed out in "Basic Concepts," the preferred method of testing writing has changed over time. Evaluating or rating of writing samples, often referred to as direct assessment, was typical procedure until about the 1940s, but ideas about writing assessment began to change. In the 1950s and 1960s, multiple choice tests of knowledge about writing, often referred to as indirect assessment, came into favor, thanks to the structuralist-psychometric ideas popular then (p. 7). But the 1970s saw a return to "language as communication," so that writing was once again assessed via rating of writing samples (p. 9). As one result, in 1986, the Test of Written English (TWE) was included on the Test of English as a Foreign Language (TOEFL), and was intended to be a direct measure of writing.

Why have objective, multiple choice writing tests not been regarded as appropriate tests of "language as communication"? Hamp-Lyons (1991) said that she did not believe that the skills needed on such tests "represent what proficient writers do" (p. 7). Similarly, Kroll (1998) said that "few in the teaching community feel

K.K. Takagi (✉)
The University of Fukui, Famille-Ai 3-201, 4418 Oyama-Cho, Machida, Tokyo 194-0212, Japan
e-mail: kjktakagi@hotmail.com

Q. Zhang (ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings*,
DOI 10.1007/978-981-10-1687-5_22

comfortable making credible claims about writers' skills on the basis of any sort of indirect measure on its own" (p. 221). Stansfield and Ross (1988) also discussed the discomfort expressed by some writing scholars over multiple choice "indirect" measures of writing. Reasons for not using objective tests of writing focus on what feels "right" in assessing writing ability, but statistical evidence for these beliefs and feelings is much harder to come by. Nevertheless, the result of discomfort with objective writing measures is that the usual method of testing writing since the 1980s, especially for placement purposes, has involved obtaining and evaluating a writing sample from students.

Rater assessment of writing samples, then, has generally been the preferred method of testing writing, but is it without problems? Not at all. There are a variety of problems associated with rating essays. First of all, the time involved can be considerable, a fact well known to writing teachers and administrators. Another difficulty is the choice of rubric or rating scale. While raters may work faster with holistic scales, the analytic rating would tend to be generally more reliable because it is comprised of a number of scores, instead of only one.

After English program administrators decide which type of scale to use, they still must choose from an assortment of rubrics, such as the TWE rubric, the Constructed Response Rubrics created by the makers of the Comprehensive English Language Test (CELT), Brown and Bailey's (1984) analytic scale, the ESL Composition Profile (Jacobs et al. 1981), and many other less known rubrics created by a wide variety of English language programs. Some of these have been evaluated and validated, but most have not. Clearly, there is a need to examine the rubrics; as Davidson (1991) pointed out, there can be serious problems in the calibration of rating scales. In examining a rubric used for rating essays for a high-stakes Japanese university entrance examination, for example, I found that the rubric favored for many years by the administrator was problematic (Takagi 2014). For example, all levels of rating categories were not used by the raters, and the threshold calibrations were too close together at three of five points. As Bond and Fox (2007) said, such small differences between steps indicate that each step was not clear in defining "a distinct position in the variable" (p. 224). Raters were not able to use the rubric completely or with precision.

Another potential problem with rating of writing samples is lack of rater agreement. If traditional methods are used to analyze results, strong interrater reliability is necessary; therefore, raters need training and practice (for a description of rater training for the TWE, see Stansfield and Ross 1988, p. 177). However, there tends to be little time available for such preparation; as a result, raters often are not experienced or trained, and, not surprisingly, their ratings of the same compositions frequently differ. In addition, raters can have individual problems in being too "safe" in using the rating scale, and therefore, overly predictable, or much worse for the measurement, unpredictable, and inconsistent.

When essays are being assessed for class placement, this problem of rater disagreement requires a procedure to resolve serious discrepancies. Some programs

follow a recommended procedure of asking a senior rater to make the final decision (Brown and Bailey 1984), but, because senior raters are not necessarily the best raters, this solution is not without problems (Takagi 2014). Other programs do not address rater disagreement at all, and simply average or total ratings. In conclusion, then, the often preferred "direct" method of rating compositions for writing assessment and placement can be fraught with difficulties, and therefore prone to inconsistencies and error.

Given the many difficulties associated with "direct" rating of compositions, surely objective writing tests can be useful tools for writing placement. Even Hamp-Lyons and Kroll admit their value. Hamp-Lyons (1991) said that these tests have correlated "fairly highly with measured writing ability" (p. 7), and Kroll (1998) said that these tests have been "valid predictors of writing ability as measured by their correlation with actual writing samples" (p. 221). Even vocal opponents to objective writing tests recognize the evidence for using them as tests of writing ability. In addition, the supposedly clear distinction between "direct" and "indirect" writing assessment is arbitrary; as McNamara (2006) said, testing is "a procedure for drawing inferences about the unobservable; it is necessarily indirect and uncertain" (p. 32). In other words, all testing is indirect in that we are attempting to measure an unobservable and latent variable, such as writing ability. Rather than creating such arbitrary distinctions between types of writing assessment, we should aim to create and validate the best writing tests possible, tests that include objective measures of writing knowledge.

In line with this aim of creating a useful objective measure of writing knowledge, I developed and pilot-tested the Sentence Form Test (SFT). There have been predecessors to this kind of writing test. Brown (1996) described the ESL placement test used at the time at the University of Hawaii as having two parts, a Writing Sample (composition), and a multiple choice proofreading test called The Academic Writing Test (p. 283). In addition, the Structure and Written Expression (Section 2) of the TOEFL, still included in the TOEFL PBT (paper-based test), is also believed to be a useful objective measure of writing skill; the Educational Testing Service (ETS) claims that it "measures the ability to recognize language appropriate for standard written English" (ETS 2016). According to Stansfield and Ross (1988), structure and written expression scores have correlated at about 0.70 with the TWE (p. 164); in other words, this objective measure of writing had a strong relationship to ratings of writing samples.

The SFT also could be called a proofreading test, but it is more precisely a test of sentence form, which tests ability to recognize correct versions of the four traditional types of sentences (simple, compound, complex, and compound-complex), the building blocks of all English writing. For each test item, students were asked to find the one incorrect sentence out of four choices. Incorrect sentences all had a serious structural error (often called major error), such as subject-verb agreement error, fragment, comma splice, etc. Such errors indicate an insufficient grasp of the language; therefore, the test was designed with the assumption that students who

recognize major errors on the SFT have a more complete knowledge of English sentences and of English writing than those who cannot do so. The test was purposely timed because, as Ellis and Barkhuizen (2005) explained, such tasks tap into implicit knowledge more than untimed tasks. In addition, the SFT was designed to tap into and account for learners' implicit knowledge, the main goal of SLA research, according to Ellis and Barkhuizen (2005). They noted that grammaticality judgment tests are very useful for "investigating specific grammatical structures that often prove difficult, or even impossible, to elicit in learner production" (p. 20). Some structures on the SFT are not usually produced by students who use English as a second or foreign language, so the SFT should work well in evaluating their knowledge of these structures.

In conclusion, then, the SFT was designed to test knowledge of English sentence structure, and it was hypothesized that this knowledge would be closely related to knowledge of English writing (as measured by performance on a writing task). As already noted, similar "indirect" tests did correlate well with ratings of writing samples; therefore, it is hypothesized that the SFT will also do so, and therefore tap into the same construct of writing ability that a composition task taps.

If the SFT and other tests like it can be shown to tap into the same construct (of writing ability) that a composition task does, then writing programs could have more options regarding writing placement test administration. For example, programs in which time and personnel abound could add another measure; multiple measures would make the writing assessment more reliable. On the other hand, if programs had no writing placement, little time, or a large number of students, then a test like the SFT alone could be used for writing placement purposes. Therefore, the specific purpose of this paper is to evaluate the SFT, especially in relation to the essay ratings given concurrently, in order to: (a) evaluate the SFT as a test and (b) determine to what extent the SFT taps into the same writing ability construct that a composition task does.

Research Questions

It is hypothesized that the SFT will work well as a writing placement test for a university EFL or ESL program, and that it will tap into the same construct of writing ability that the writing section of the test taps. In order to test this hypothesis, the following research questions were posed:

Research Question 1: Does the SFT work well as a writing test in that test items match student ability, create a useful spread of student ability, display acceptable fit values for the Rasch model, demonstrate acceptable reliability, and are unidimensional in measuring one construct?
Research Question 2: Can the SFT be shown to tap into same construct of writing ability that a composition task taps?

Method

Participants

Fifty freshman students at a women's college in Tokyo, Japan, took a writing placement test after finishing one year of composition study. Forty-five students were Japanese, and five were Chinese exchange students, all approximately 19 years of age. At the time the writing placement test was administered, students had all studied English for approximately seven years: six years in junior high and high school, and one year in college. Most Japanese students had not studied or lived overseas. Their language proficiency varied from basic to high intermediate; specifically, their Pre-TOEFL ITP scores from January of the same year ranged from a low of 293 to a high of 480 (mean of 383.20 and standard deviation of 36.40). Results of the writing placement test were to be used to place students into second-year writing classes.

Materials

The writing placement test included two sections. The first section was a composition in which students were asked to write about what they had learned in their first-year at university. Since instruction in the five levels of the first-year writing classes varied considerably, students were allowed to write either an essay or a long paragraph of about 250 words. The time limit for the writing assignment was 40 min. The second section of the test was the Sentence Form Test (SFT). On each test item, they were asked to find one incorrect option out of four example sentences. Students were given 20 min to complete the test. The following are directions and an example item included on the test:

Read the four sentences for each question. One of the sentences has an important mistake. Write the letter of the sentence with the mistake.

EXAMPLE

(a) I am late.
(b) I late.
(c) I was late.
(d) I was not late.

Answer: b

Scoring

The SFT answer sheets were quickly scored (in about 15 min). The 50 compositions were scored by four raters, all university teachers in Japan. Three of the raters

used Brown and Bailey's (1984) analytic scale, and two raters used a holistic scale used by the English program at Hawaii Pacific University. The analytic raters were all university EFL composition teachers; one was American, the second was British, and the third was Chilean. The holistic raters were Americans teaching college EFL. The process of scoring the essays took approximately four hours. The analytic scale yielded a maximum of 100 (with a maximum of 20 each for (a) organization; (b) logical development of ideas; (c) grammar; (d) punctuation, spelling, and mechanics; and (e) style and quality of expression). The holistic scale was originally based on a 0–10 point scale but was converted to a 100-point scale.

Procedures and Data Analysis

In order to answer the research questions, a number of statistical analyses were used. In answering the first question (determining whether the SFT generally worked well as a writing placement test), I conducted a Rasch analysis using Winsteps, version 3.90.0, in order to examine the variable map, and the fit and difficulty of SFT items. I examined test reliability with the Rasch model, as well as through traditional methods for assessing internal consistency. I also investigated unidimensionality of the SFT by examining a bubble chart pathway plot produced with Winsteps.

In answering the second research question (determining whether the SFT taps into the construct of writing ability tapped by writing task ratings) I first examined intercorrelations of SFT scores and writing task ratings because correlation coefficients indicate the degree to which measures "tap the same construct" (Stansfield and Ross 1988, p. 168). I then examined unidimensionality through principal components analysis.

Results

Research Question 1: Does the SFT work well as a writing test in that test items match student ability, create a useful spread of student ability, display acceptable fit and difficulty, demonstrate acceptable reliability, and are unidimensional in measuring one construct?

Descriptive statistics for essay ratings and test scores are shown in Table 1.

Figure 1 shows a variable map for the SFT produced by Rasch Analysis. The software used was Winsteps, Version 3.90.0, developed by J.M. Linacre. The 20 test items are on the right side, with most difficult (item 16) at the top, and least difficult at the bottom. The 50 students are on the left. The student with highest ability (student 22) is at the top, and least able students are shown at the bottom. The variable map shows that test items are mostly a good match for the students,

Table 1 Mean scores and statistics for essay ratings and SFT

Measure	HR1	HR2	AR1	AR2	AR3	SFT
M	48.40	59.96	68.52	67.34	55.94	9.44
SD	17.77	14.70	10.18	10.21	15.90	3.91
Skewness	0.50	0.14	−0.34	0.03	−0.21	0.41
SE of skewness	0.34	0.34	0.34	0.34	0.34	0.34
Kurtosis	−0.17	−0.29	0.16	−1.01	−0.52	−0.44
SE of kurtosis	0.66	0.66	0.66	0.66	0.66	0.66

Note $N = 50$ for all ratings. *HR1* holistic rater 1 scores; *HR2* holistic rater 2 scores. *AR1* analytic rater 1 scores; *AR2* analytic rater 2 scores; *AR3* analytic rater 3 scores. *SFT* Sentence form test scores. Possible score range for essay ratings is 0–100, and for SFT, 0–20 points

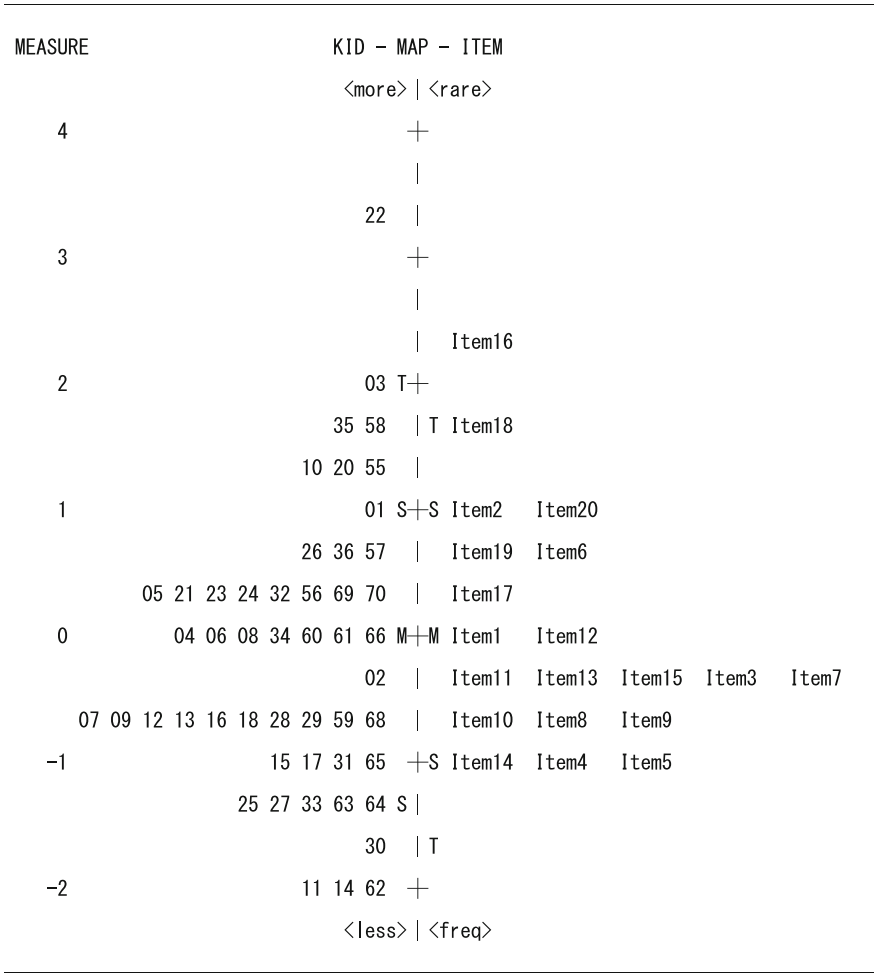

Fig. 1 Variable map for person ability and item difficulty of the SFT

and there is a good spread of item difficulty. Ability levels of students are spread out in a way that is useful for placement into writing classes.

The fit and difficulty of test items were assessed using the Rasch model. Table 2 shows the Infit Mean Square (Infit MNSQ), Outfit Mean Square (Outfit MNSQ), and Measure (indicating difficulty) for each item. According to Bond and Fox (2007), mean square infit and outfit values for a multiple choice high-stakes test should range from 0.8 to 1.2, and for a "run of the mill" multiple choice test, from 0.7 to 1.3 (p. 243). These "run of the mill" values would be acceptable for a writing class placement test.

Table 2 shows (in the Measure column) that the items generally move from easier to more difficult items, as was intended in the test design. However, this progression is not perfect, and some items do not follow the intended pattern. For example, items 2 and 6 are not as easy as later items, while item 14 is easier than intended. The Infit and Outfit MNSQ columns show that almost all test items fit the model well, though both infit and outfit values for Item 20 are too high.

Results from the Rasch Analysis revealed that the SFT item reliability was 0.85, and the student reliability was 0.73. Other traditional methods for assessing internal consistency were also used, for purposes of comparison. As recommended by Brown (1996, pp. 194–203), the split-half method adjusted by using the Spearman-Brown prophecy formula was employed. Since the test was designed to be progressively

Table 2 Rasch model descriptors of SFT test items

Items	Infit MNSQ	Outfit MNSQ	Measure
Item 1	0.89	0.83	−0.09
Item 2	0.86	0.75	0.86
Item 3	1.01	0.92	−0.19
Item 4	0.85	0.76	−1.09
Item 5	0.96	0.97	−0.99
Item 6	1.17	1.38	0.63
Item 7	0.84	0.79	−0.38
Item 8	1.04	0.99	−0.68
Item 9	0.94	0.93	−0.68
Item 10	0.73	0.64	−0.78
Item 11	1.07	1.07	−0.48
Item 12	0.94	0.96	−0.09
Item 13	1.09	1.07	−0.19
Item 14	0.84	0.73	−1.09
Item 15	1.02	0.93	−0.38
Item 16	0.87	0.65	2.24
Item 17	1.17	1.41	0.21
Item 18	1.28	1.30	1.67
Item 19	1.08	1.21	0.52
Item 20	1.34	1.61	0.97

Note $N = 50$

more difficult, splitting it into two halves by dividing odd-numbered from even-numbered items made sense. The resulting Spearman-Brown Coefficient was 0.82. Other coefficients were also produced. Cronbach's Alpha for Part 1 was 0.58, and for Part 2, 0.63; the correlation between forms was 0.69. Finally, the Guttman Split-Half Coefficient was 0.82. In conclusion, then, the SFT could be considered reliable for this group of students. Perhaps it could also be reliably used with a group of students with similar English proficiency (students who score from approximately 290–500 on the Institutional Pre-TOEFL).

I also examined unidimensionality of SFT test items through inspection of a bubble chart pathway plot produced by Rasch Analysis, using Winsteps. Figure 2 shows the plot; it is a representation of the fit of items of the SFT (specifically looking at infit). According to Bond and Fox (2007), fit statistics help us "to determine whether the item estimations may be held as meaningful quantitative summaries of the observations (i.e., whether each item contributes to the measurement of only one construct)" (p. 35). Acceptable fit on this plot is between −2 and +2 (p. 57). In addition, the size of the circles in this plot is a reflection of error, with larger circles reflecting more error in measurement. In the figure below, the results are generally positive regarding the fit of SFT items because most test items fit within the range of −2 and +2. However, we also can see that some items may need revision; item 10 is overfitting while item 20 is close to underfitting. In addition, items 16 and 18 display relatively more error and should be examined for possible revision as well. In short, despite the need to inspect a few items, this line

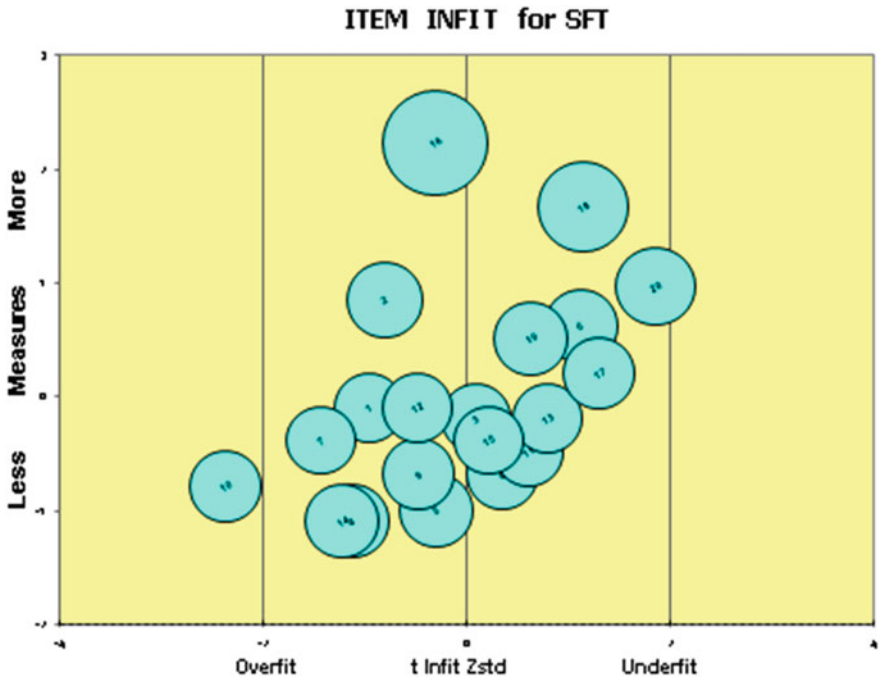

Fig. 2 Item infit for SFT items

of validation evidence for the use of the SFT is mostly positive in that almost all items are contributing in a meaningful manner to measurement of one construct.

The results of Research Question 1 are positive validation evidence in that the test items match student ability, create a useful spread of student ability, generally display acceptable fit and difficulty, and demonstrate acceptable reliability as well as unidimensionality. The test content is achieving its purpose, and generally working in a positive way to measure students in a unidimensional manner.

Research Question 2: Can the SFT be shown to tap into same construct of writing ability that a composition task taps?

In order to answer research question 2, I first examined correlations among all scores. The Pearson product-moment correlation coefficient was calculated for all the comparisons. Specifically, the SFT scores were correlated with the essay scores (produced by two holistic raters and three analytic raters). Naturally, the assumptions underlying the correlation statistic (of independence, normal distribution, and linear relationship) were checked (Brown 1996, p. 157). All assumptions were met. The unadjusted correlations are presented in Table 3. The Bonferroni approach was used to control for Type I error across the 15 correlations. A p value of less than 0.003 (0.05/15 = 0.003) was required for statistical significance (Green and Salkind 2005, p. 261). The results showed that all 15 coefficients were statistically significant and large (Field 2005). Such results suggest that students tended to score in a similar fashion on the SFT and the writing task. As Stansfield and Ross (1988) said, this result suggests that the SFT and writing task both tap into the same construct of writing ability.

I also examined the SFT and essay ratings for unidimensionality using principal components analysis. This type of analysis allows us to examine underlying dimensions, and to determine whether test scores "reflect a single variable" or not (Field 2005, p. 619). The analysis was conducted in order to determine whether the SFT would load together with essay ratings onto the same factor. Table 4 presents the results. The analysis resulted in one component, and an eigenvalue of 3.62, accounting for 72.24 % of the variance. According to Armor (1974), any factor that accounts for 40 to 60 % is a good solution; therefore, these results are favorable. In short, the essay ratings and SFT were fundamentally unidimensional, and seem to

Table 3 Intercorrelations of holistic essay ratings, analytic essay ratings, and SFT scores

Measure	1	2	3	4	5	6
1. HR1	–					
2. HR2	0.83[a]	–				
3. AR1	0.71[a]	0.60[a]	–			
4. AR2	0.69[a]	0.57[a]	0.72[a]	–		
5. AR3	0.70[a]	0.56[a]	0.82[a]	0.82[a]	–	
6. SFT	0.65[a]	0.57[a]	0.63[a]	0.57[a]	0.55[a]	–

Note [a]$p < 0.0001$. $N = 50$ for Essay Ratings and SFT. *HR* Holistic rater; *AR* Analytic rater. *SFT* Sentence form test

Table 4 Factor loadings from principal components analysis of essay ratings and SFT: communalities, eigenvalue, and percentage of variance

Writing measure	Component 1	Communality
HR1	0.90	0.81
HR2	0.81	0.65
AR1	0.88	0.77
AR2	0.86	0.74
AR3	0.88	0.77
SFT	0.77	0.60
% of variance	72.24	

be tapping into the same construct of writing ability. Although a larger sample size would be preferable for principal components analysis, this line of validation evidence also supports using the SFT as a test of writing.

Discussion

As an objective measure of writing ability, the SFT has many obvious advantages for writing class placement. It is administered and scored quickly and easily, and there are no concerns about choice or quality of rubrics, or about rater behavior. Although some may argue the need for "direct" writing measures, because these feel somehow "right," surely professionals must use more than feelings in making testing decisions. Though ratings of writing tasks can work well, as they did in this study, it is clear that good objective measures like the SFT can offer an efficient and reliable supplement to or substitute for traditional rater assessment of writing.

References

Armor, D. J. (1974). Theta reliability and factor scaling. *Sociological Methodology, 5*, 17–50.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, N.J.: Lawrence Erlbaum.

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

Brown, J. D., & Bailey, K. M. (1984). A categorical instrument for scoring second language writing skills. *Language Learning, 34*, 21–42.

Davidson, F. (1991). Statistical support for training in ESL composition rating. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 155–164). Norwood: Ablex Publishing.

Ellis, R., & Barkhuizen, G. (2005). *Analysing learner language*. Oxford: Oxford University Press.

ETS. (2016). Retrieved February 29, 2016, from: https://www.ets.org/toefl/pbt/prepare/sample_questions/structure_written_expression_practice_section2

Field, A. (2005). *Discovering statistics using SPSS* (2nd ed.). London: Sage.

Green, S., & Salkind, N. (2005). *Using SPSS for Windows and Macintosh: Analyzing and understanding data* (4th ed.). Upper Saddle River, New Jersey: Pearson Prentice Hall.

Hamp-Lyons, L. (1991). Issues and directions in assessing second language writing in academic contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 323–329). Norwood: Ablex Publishing.

Jacobs, H. L., Zinkgraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition: A practical approach*. Rowley: Newbury House Publishers.

Kroll, B. (1998). Assessing writing abilities. *Annual Review of Applied Linguistics, 18*, 219–239.

McNamara, T. (2006). Validity in language testing: The challenge of Sam Messick's Legacy. *Language Assessment Quarterly, 3*(1), 31–51.

Stansfield, C. W., & Ross, J. (1988). A long-term research agenda for the test of written english. *Language Testing, 5*, 160–186.

Takagi, K. K. (2014, August). *Writing assessment in university entrance examinations: The case of one Japanese university*. Paper presented at the meeting of the Pacific Rim Objective Measurement Symposium, Guangzhou, China.