Language Testing in Asia
a SpringerOpen Journal

RESEARCH Open Access

Rater reliability and score discrepancy under holistic and analytic scoring of second language writing

Bo Zhang[1*], Yunnan Xiao[2] and Juan Luo[2]

* Correspondence: boz@uwm.edu
[1]Department of Educational Psychology University of Wisconsin – Milwaukee, P O Box 413, Milwaukee WI 53201-0413, USA
Full list of author information is available at the end of the article

Abstract

Previous studies comparing holistic scoring to analytic scoring of second language writing have given mixed results. Some of them suffer from methodological drawbacks, such as limited writing sample size, limited number of raters, and lack of direct comparison of the two methods. Based on 300 writing samples graded by 14 raters, this research continues the comparison of the two scoring methods in two ways: examine rater reliability for each method and investigate the discrepancy of the scores assigned by them. Results show while rater reliability is quite high and similar for the two methods when a large number of raters are used, the scores assigned can be quite different. Specifically, students with lower writing proficiency tend to receive higher scores under analytic scoring while students with higher proficiency score higher under holistic scoring.

Keywords: Rater reliability; Holistic scoring; Analytic scoring; Generalizability theory; Second language writing

In language testing, the debate between holistic and analytic scoring of writing tasks has been long and well-documented. The focus of the debate is on which method is able to provide more valid and reliable scores in measuring writing skills. The proponents of the holistic approach have suggested that writing be scored impressionistically and a holistic reading of an essay should involve reading for an individual impression of the quality of the writing (e.g. Cooper and Odell 1977). According to this approach, the construct of writing should be treated as a single entity that integrates the inherent quality of writing and that quality can be recognized only by experienced readers using skilled impression (White 1984). When raters are well trained, the reliability by holistic scoring can be very high (Cooper and Odell 1977).

On the other hand, previous research has also demonstrated that applying the same criteria in holistic scoring is not an easy task by multiple raters, hence, the application of the holistic scoring is likely to engender unreliable ratings (Diederich et al. 1961). Holistic scores have also been shown to correlate with relatively superficial characteristics of writing, such as text length and handwriting style (Fulcher 1997; Markham 1976). To increase rater reliability, a scoring rubric should be written clearly so that different raters will be able to use the same qualities inherent to the text in assigning their scores (Diederich et al. 1961). This last statement actually puts holistic scoring more in line with analytic scoring.

In theory, analytic scoring is less susceptible to rater deviance. This method quantifies multiple aspects of a task and scores them separately. For writing, features such as content, organization, cohesion, vocabulary, grammar, and mechanics are often meaningful and thus scored. Many researchers believe (e.g., Klein et al. 1998) that analytic scoring provides a more objective assessment of writing quality than the holistic method. In addition, as scores on multiple aspects of writing are assigned, analytic scoring can give more diagnostic information to second language learners, who usually have an uneven profile across different aspects of writing (Hamp-Lyons 1991; Johnson and Hamp-Lyons 1995). On the other hand, research has also revealed the disadvantages in analytic scoring. Underhill (1987) has shown that it is almost impossible for a rater to keep track of more than three features simultaneously. Consequently, raters usually focus on one feature at a time. This repetitive work can lead to mental fatigue in scoring. Holistic scoring, on the other hand, tends to alleviate these cognitive burdens (Douglas and Smith 1997). Bauer (1981) showed that analytic scoring takes twice as long to train raters and four times as long to grade than holistic scoring.

How these two methods compare has been investigated extensively. Overall, findings have been mixed. While some studies found high correlation (e.g., Bauer 1981; Vacc 1989) or high level of similarity between these two methods (e.g., Bacha 2001), other studies have given advantage to one or the other (e.g., Veal and Hudson 1983; Swartz et al. 1999; Nakamura 2002; Schoonen 2005). Many studies on this topic use correlation to evaluate rater reliability. As correlation by nature reflects the relative ranking of subjects on two measures, findings from those studies can only be applied to the so-called relative decision making or norm-referenced score interpretation. In reality, important decisions are often made based on the absolute value of individual scores. In writing, determining whether a student can write or not is as important as, if not more than judging whether a student can write better than his or her peers. Another limitation in previous studies is that a rating scale is not clearly defined or consistently used. In some cases, writing sample size and number of raters are both small.

This study investigates the scoring reliability by using the generalizability theory analysis (G theory) (Shavelson and Webb 1991). The advantages of using the G theory to assess rater reliability have been clearly explained in Swartz et al. (1999) and Zhang et al. (2008). In essence, G theory empowers researchers to decompose the measurement error in a study into multiple meaningful components, which in turn, can be incorporated into the decision making process. In grading writing tasks, one major factor (known as facet in the G theory terminology) is rater effect. Rater effect indicates the variability in scores due to raters. For instance, some raters are stricter in their ratings than others. Facets can be crossed or nested in a G-study. If all raters grade all writings, the rater facet will be crossed with the person facet. On the other hand, if different raters grade different persons, the person effect will be nested within the rater effect.

Swartz et al. (1999) evaluated the scoring reliability in assessing writing for the first language. It demonstrated a lack of reliability in both holistic and analytic scorings in many conditions when .9 was set as the criterion for acceptable high level of reliability. Akin to the Swartz et al. (1999), G-theory is applied in the current study. Unlike that study, this study focuses on second language writing. Specifically, the study aims to provide a clear picture of how scores under the holistic and analytic methods compare by using a large writing sample and a large number of raters. The comparison will be conducted in two ways. First, scores under each scoring method will be compared to determine whether raters are able to provide similar scores for each sample. Second, scores will be compared across the two methods to see whether similar scores are assigned to each subject under two different scoring methods. Formally, the following three research questions will be addressed:

1. How does rater reliability under holistic and analytic scoring compare in grading second language writing?
2. Under analytic scoring, is rater reliability similar across different writing components?
3. Do students receive similar scores under holistic and analytic scoring methods?

Method

Design

This study used a fully crossed design. As illustrated in Figure 1, the rater facet is crossed with the person facet. In practice, this indicates that each rater grades each person twice, once by the holistic method and the second time by the analytic. Using the G-theory terminology, the variance components under this design are: persons (σ_p^2), raters (σ_r^2), and person by rater interaction which is inseparable from the residual error ($\sigma_{pr,e}^2$). The total score variance is decomposed into three components, or

$$\sigma^2(X_p) = \sigma_p^2 + \sigma_r^2 + \sigma_{pr,e}^2 \tag{1}$$

The higher the proportion of the total variance accounted for by the person variance, the higher the reliability. In other words, the writing score will be reliable if score difference reflects the difference in writing skills of students rather than the difference of how each rater has graded.

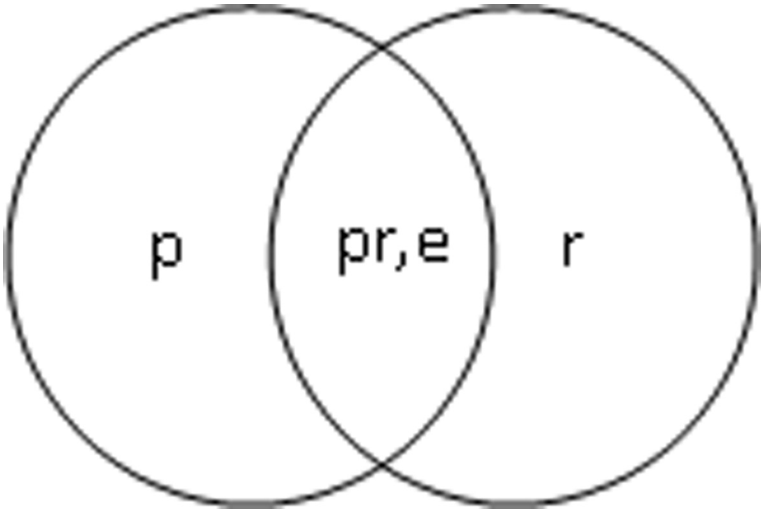

Figure 1 G-study design for L2 writing grading.

The generalizability coefficient ($\mathrm{E}\rho^2$) addresses the relative error in making decisions. It is computed as:

$$\mathrm{E}\rho^2 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{rel}^2} = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{pr,e}^2}{n_r}} \tag{2}$$

where σ_{rel}^2 is the relative error; n_r is the number of raters, and other terms share the same meaning as in Equation 1. Meanwhile, the index of dependability (ϕ) (Brennan & Kane, 1977) measures the consistency in making an absolute decision, which focuses on the absolute writing level of an individual with no reference to other students. The index of dependability is expressed as:

$$\phi = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{abs}^2} = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_r^2}{n_r} + \frac{\sigma_{pr,e}^2}{n_{rt}}} \tag{3}$$

where σ_{abs}^2 is absolute error and all other terms have been defined above.

Samples and raters

The task was a writing task from a college English placement test for the incoming freshmen in a major Chinese university. Altogether over 5,000 students were tested and their writing was graded by professors and instructors in the English teaching program. For this project, 300 writing samples were selected from that large group by stratified sampling. Specifically, for each level from 1 to 10, 30 samples were selected. These samples were then rated by 14 raters using the two grading methods.

The raters were graduate students in a three-year language testing program. The length of their time in the program is as follows: 3 in the third year, 4 in the second year, and 7 in the first year. All raters had English teaching experience. They also had participated in at least two writing grading projects in the past.

Scoring

Rating scales were developed with reference to the Centre for Canadian Language Benchmarks (2000). The scales were composed of 11 levels. The following five components were scored: task, grammar, mechanics, vocabulary, and structure. Under the holistic scale, raters simultaneously took all these five components into consideration and assigned each student to one level. Under the analytic scoring scale, raters scored each component first and then added the scores up to derive an overall level. As very few students scored 11, they were not part of the sample.

The following steps were taken in rater training. Raters first studied the two scoring rubrics. They were then given 10 anchor samples to grade. After that, they compared their ratings to the actual level of each sample. Finally, they discussed their ratings as a group. Scoring of all 300 samples took one and half days under holistic scoring and another eight and half days under analytic scoring. To reduce fatigue and possible memory effect, there was a five-day break between the holistic and analytic grading periods.

Results

Results will be presented in the order of the above three research questions. Recall that the first question is on rater reliability. Table 1 presents variance components under the two scoring methods.

As shown in the table, under both holistic and analytic scoring frames, person facet, or student writing level, accounted for the largest variance in the writing scores. Rater variance was small and similar across the two methods, accounting for less than 5% of the total variance. The magnitude of the error variance was also similar for the two methods.

Next, the reliability for the absolute and relative decisions is presented in Table 2. Under the G-theory framework, one may not only examine the reliability under the current design (i.e., 14 raters grading 300 samples), reliability in other interesting conditions may also be predicted once the variance component for each facet is estimated. Thus in the table, the reliability for the typical conditions in writing assessment are presented along with the current condition.

In general, the analytic holistic scoring enjoyed a slim .01 to .04 edge over the holistic way. Reliability for the relative decision was slightly higher (.01 to .03) than that for the absolute decision, as expected. In one of the most common conditions where each classroom teacher (i.e., one rater) grades each student writing, reliability would be in the .8 range, which should be acceptable given that teachers usually give multiple assessments throughout a semester or writing is usually part of a language test. But using the standard of .9 set in Swartz et al. (1999) for high-stake decisions, such as the placement test in the current study, two raters would be required for grading each student.

The second research question asks whether rating by the components in the analytic scoring is equally reliable. As shown in Table 3, rater reliability was lower for the component scores than that for the overall level. Using the .9 cut-off value would require 4 raters being used. On the other hand, the reliability difference among these components was quite small. Note that for both relative and absolute decisions, the "structure" component had the lowest rating reliability.

The final research question is on the discrepancy of scores assigned by the two methods. The overall distribution of the two scores looked quite similar: for holistic scoring: $M = 5.03$ and $SD = 1.79$; for analytic scoring: $M = 5.01$ and $SD = 1.74$. A paired *t*-test showed no significant difference between them, $t(299) = 0.15$, p=.88. However, the analysis of the difference score between the holistic and analytical scores painted a different picture, as illustrated in Figure 2. The difference score in the graph is computed by simply subtracting the analytic score from the holistic score, thus a value of 1 in the graph indicates that the rating by the holistic method is one level higher than that by the analytic method. When the group of 300 subjects was taken as a whole, the

Table 1 Variance components and percentage

Effect	Holistic scoring		Analytic scoring	
	Variance component	% of Total variance	Variance component	% of Total variance
Person	2.53	78	2.92	82
Rater	0.13	4	0.05	2
Error	0.56	18	0.50	17

Table 2 Reliability as a function of number of raters

No. of raters	Holistic scoring		Analytic scoring	
	Relative	Absolute	Relative	Absolute
1	0.82	0.78	0.83	0.82
2	0.90	0.88	0.91	0.90
3	0.93	0.92	0.94	0.93
14	0.98	0.98	0.99	0.98

This is the condition in the present study.
Note. The last row with 14 raters shows the results for the condition in the present study

distribution of the difference score was approximately normal, indicating there was about equal number of higher ratings and lower ratings under each method.

Next, to address such practical questions as "what percentage of ratings is different by 1 level under these two methods?" the differences in the above figure were aggregated regardless of the direction of difference. Out of the 4,200 pairs (300 samples graded by 14 raters), 21.7% had perfect agreement. If one point deviation was tolerated, the agreement level rose to 54.8%. If two points deviation was allowed, the agreement increased further to 74.7%. However, 12.6% pairs were off by more than 3 levels and 1.7% or 71 pairs were off by more than 5 levels, which would be unacceptable in any situation.

Finally, how the two scoring methods affect students with different writing levels was studied. In Figure 3, values on the horizontal axis are the original values assigned by professors or lead instructors, thus treated as a proxy to the actual level of each student's writing level. Values on the vertical axis were the difference between the holistic and analytic scores, which shares the same meaning as in Figure 2. Interestingly, there is a clear pattern. For students with low writing levels (i.e., levels 1, 2, and 3), on average, their holistic scores were one to two levels lower than their analytic scores. For example, for level 1, the average holistic score was 2.33 levels lower. For students with medium writing levels (i.e., levels 4, 5, and 6), the difference between the two methods was less than one level. Take level 5 as an example. The average difference was 0. In other words, on average, the two methods were able to provide exactly the same score for these students. For students with relatively high writing levels (i.e., 7, 8, 9 and 10), their holistic scores were higher by one to one and a half levels.

Table 3 Reliability for the components under analytic scoring

Decision to be made	No. of raters	Task	Grammar	Mechanics	Vocabulary	Structure
Relative	1	0.77	0.73	0.72	0.75	0.70
	2	0.87	0.85	0.84	0.85	0.83
	3	0.91	0.89	0.88	0.90	0.88
	4	0.93	0.92	0.91	0.92	0.90
Absolute	1	0.72	0.72	0.70	0.73	0.69
	2	0.84	0.83	0.82	0.84	0.82
	3	0.89	0.88	0.87	0.89	0.87
	4	0.91	0.91	0.90	0.91	0.90

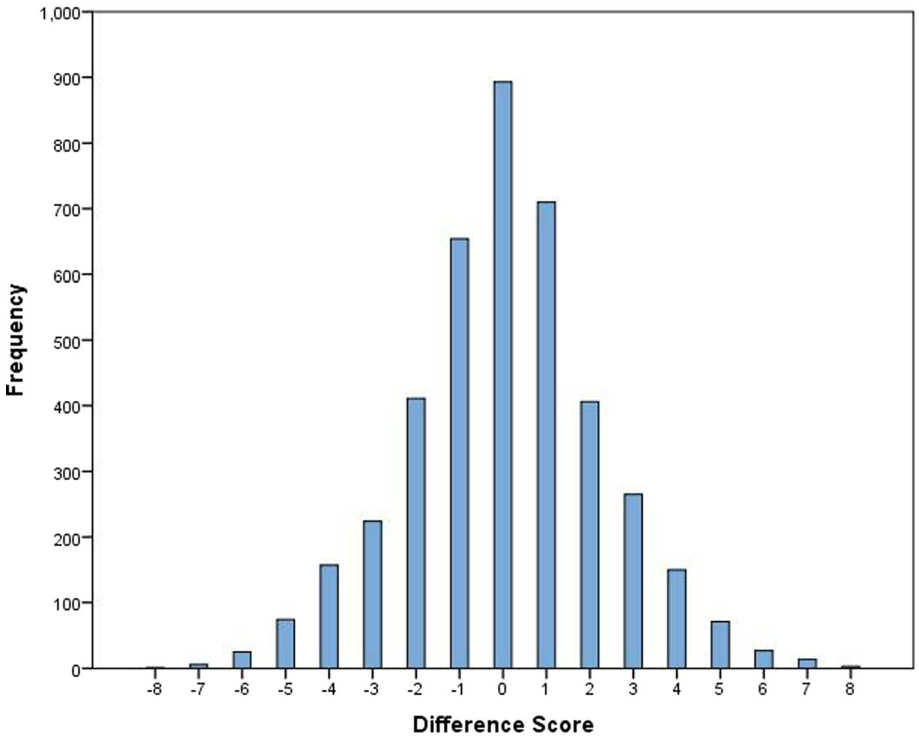

Figure 2 Distribution of the difference score between holistic and analytic scoring.

Discussion

This study was set up to investigate rater performance under holistic and analytic scorings by using a large number of writing samples and a large number of raters. Findings from this study support the notion that high rater reliability can be achieved under each method. On the other hand, this study also reveals that scores assigned under the two methods can be very different.

For everyday classroom instruction, using one rater would give the rater reliability around .8 by either method, which should be high enough, given that writing is usually part of an overall language proficiency assessment that includes other components such

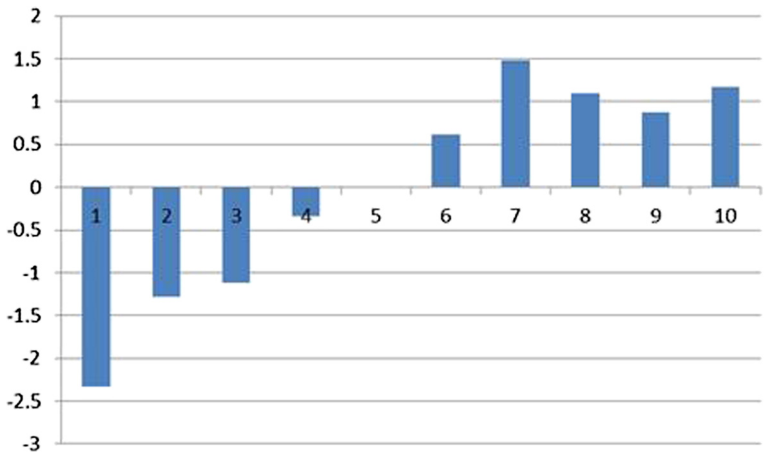

Figure 3 Distribution of the difference score by the writing level.

as reading and grammar. However, achieving this high reliability by one rater requires raters being well trained. Recall that raters in this study were trained in language testing, studied the scoring rubrics carefully, and graded pilot samples before conducting the actual grading. In other words, these raters understood second language writing, knew what to focus on in their grading, and were calibrated by pilot samples. If any of these components is missing, reliability is likely to suffer and more raters may be necessary.

In analytic scoring, rating by components shows lower reliability than that by the overall level. This may be due to the fact that while it is not hard to judge the overall writing level, to decide a specific level (e.g. a 4 instead of a 5) for one component of the writing is more challenging. Still, rater reliability is satisfactory, which implies that component scores may be provided to describe the strengths and weaknesses in writing. Actually, that is the major advantage of analytic grading: to give more specific feedback so that students can improve their writing more effectively.

This study also reveals inconsistency in grading L2 writing under holistic and analytic methods. Scores assigned to individual students by these methods can be very different, which certainly is problematic. Moreover, that difference varies by writing level. While the current study is not able to explain this difference, one possible reason is while writing from lower level students may look pretty bad when graded as a whole, some components may look decent when graded separately under analytic scoring. On the other hand, writings from higher level students may look decent as a whole but under the microscope of analytic scoring, they may show deficiency in some components, resulting in lower scores under analytic scoring.

It is not the purpose of this study to ascertain whether holistic or analytic method should be preferred in grading L2 writing. Rather, the goal was to provide more information on how these two methods resemble or differ so that practitioners can benefit most in using either of them. With regard to which method one should use, in addition to rater reliability and score discrepancy as investigated in this study, one would also have to take into many other factors, such as the purpose of testing, rater qualification and availability, time, and cost. Take time as an example, holistic scoring of the 300 samples took 1.5 days whereas analytic scoring took 8.5 days. That difference would give a definitive edge to holistic scoring in situations when time is tight or prompt score reporting is required.

One limitation of this study lies in the representativeness of the sample. While the sample size is relatively large, as all subjects were college freshmen from one university, the sample reflects quite homogeneous writing skills. On indication of that is that no sample was selected from the highest level of writing, resulting in a possible over-representation of the lower levels of writing. As this study actually shows that holistic scoring tends to give more credit to higher levels, future research with more students at the high end would provide more evidence on this issue. Another direction for future research is to explore ways to combine the holistic and analytic scores into a mixed score, which will outperform both of them.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

BZ conducted literature review, conducted statistical analyses, drafted and revised the manuscript. YX conducted data collection, assisted with data analysis, and revised the manuscript. JL conducted literature review, assisted with data collection and revised the manuscript. All authors read and approved the final manuscript.

Authors' information

BZ is an associate professor in Department of Educational Psychology, University of Wisconsin – Milwaukee, USA. His main research interests include language assessment, large-scale assessment, and item response theory. YX is a professor of English in College of Foreign Languages, Hunan University, China. Her research areas include writing assessment, placement testing, and second language acquisition.

Acknowledgement

This study was partially supported by China National Philosophy Social Science Planning Project, approval serial No. 08BYY025.

Author details

[1]Department of Educational Psychology University of Wisconsin – Milwaukee, P O Box 413, Milwaukee WI 53201-0413, USA. [2]College of Foreign Languages, Hunan University, Changsha, Hunan 410082, PR of China.

Received: 19 December 2014 Accepted: 19 February 2015
Published online: 05 March 2015

References

Bacha, N. (2001). Writing evaluation: what can analytic versus holistic essay scoring tell us? *System, 29*(3), 371–383. doi:10.1016/S0346-251X(01)00025-2.

Bauer, B. A. (1981). A study of the reliabilities and the cost-efficiencies of three methods of assessment for writing ability. University of Illinois, Champaign, IL.

Brennan, RL, & Kane, MT. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement, 14*(3), 277–289.

Centre for Canadian Language Benchmarks. (2000). Canadian Language Benchmarks 2000 – English as a second language for adults. Center for Canadian Language Benchmarks, Ottawa, Canada.

Cooper, C, & Odell, L. (1977). Holistic evaluation of writing. In CR Cooper & L Odell (Eds.), *Evaluating Writing* (pp. 3–31). Urbana, IL: National Council of Teachers of English.

Diederich, PB, French, JW, & Carlton, ST. (1961). *Factors in judgments of writing ability* (Research Bulletin RB-61-15). Princeton, N.J: Educational Testing Service.

Douglas, D, & Smith, J. (1997). *Theoretical underpinnings of the Test of Spoken. English Revision Project. (TOEFL Monograph Series No. 9)*. Princeton, NJ: ETS.

Fulcher, G. (1997). The testing of speaking in a second language. In C Clapham & D Corson (Eds.), *Encyclopedia of language education, Vol. 7: Language testing and assessment* (pp. 75–78). Dordrecht, Netherlands: Kluwer Academic Publishers.

Hamp-Lyons, L. (1991). *Assessing Second Language Writing in Academic Contexts* (pp. 279–291). Norwood. NJ: Ablex Publishing Corporation.

Johnson, DM, & Hamp-Lyons, LIZ. (1995). Research on the Rating Process: Rating Nonnative Writing: The Trouble with Holistic Scoring. *Tesol Quarterly, 29*(4), 759–762. DOI:10.2307/3588173.

Klein, SP, Stecher, BM, Shavelson, RJ, McCaffrey, D, Ormseth, T, Bell, RM, Comfort, K, & Othman, AR. (1998). Analytic versus holistic scoring of science performance tasks. *Applied Measurement in Education, 11*, 121–137. doi:10.1207/s15324818ame1102_1.

Markham, LR. (1976). Influences of handwriting quality on teacher evaluation of written work. *American Educational Research Journal, 13*(4), 277–283. doi:10.3102/00028312013004277.

Nakamura, Y. (2002). A comparison of holistic and analytic scoring methods in the assessment of writing. The Interface between Interlanguage, Pragmatics and Assessment: Proceedings of 3rd Annual JALT Pan-SIG Conference. May 22–23, Tokyo, Japan: Tokyo Keizai University.

Schoonen, R. (2005). Generalizability of writing scores: an application of structural equation modeling. *Language Testing, 22*(1), 1–30. doi:10.1191/0265532205lt295oa.

Shavelson, RJ, & Webb, NM. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Swartz, CW, Hooper, S, Montgomery, J, Wakely, M, Renee, E, Kruif, D, & Reed, M. (1999). Using generalizability theory to estimate the reliability of writing scores derived from holistic and analytical scoring methods. *Educational and Psychological Measurement, 59*, 492–506. doi:10.1177/00131649921970008.

Underhill, N. (1987). *Testing Spoken Language: A Handbook of Oral Testing Techniques*. Cambridge: Cambridge UP.

Vacc, NN. (1989). Writing evaluation: examining four teachers' holistic and analytic scores. *The Elementary School Journal, 90*(1), 87–95.

Veal, LR, & Hudson, SA. (1983). Direct and indirect measures for large-scale evaluation of writing. *Research in the Teaching of English,* 290–296.

White, E. (1984). Holisticism. *College Composition and Communication, 35*, 400–409.

Zhang, B, Johnston, L, & Kilic, GB. (2008). Assessing the reliability of self- and peer rating in student group work. *Assessment & Evaluation in Higher Education, 33*(3), 329–340. doi:10.1080/02602930701293181.