



Title of Project:

The Effects of Primacy on Rater Cognition:
An Eye-tracking Study

Researcher:

Laura Ballard
Michigan State University
lauradballard@gmail.com



Laura Ballard

Research Supervisor:

Dr. Paula Winke
Michigan State University
winke@msu.edu

Project Summary

Motivation for the Research

Rater scoring has an impact on performance test reliability and validity. Thus, there has been a continued call for researchers to investigate issues related to rating (Crusan, 2015). Myford (2012) exhorts researchers and test designers to “do all that [they] can to help ensure that the ratings that raters assign are accurate, reliable, and fair” (p. 49). Second language testing researchers are committed to this goal and have been researching the various facets that affect test scoring processes for years (Cumming, Kantor, & Powers, 2002; Eckes, 2008; Kondo-Brown, 2002; Lumley, 2002; Orr, 2002). In second language writing assessment, such emphasis on investigating the scoring process and how raters arrive at particular scores have been seen as critical “because the score is ultimately what will be used in making decisions and inferences about writers” (Weigle, 2002, p. 108).

In the current study, I answer the call for continued research on the rating process by investigating rater cognition in the context of writing assessment. Research on raters’ cognitive processes “is concerned with the attributes of the raters that assign scores to student performances, and their mental processes in doing so” (Bejar, 2012, p. 2). A theme central to rater cognition is the way in which raters interact with rubrics. Only by understanding this interaction will test designers be able to improve rubrics, rater training, and test reliability and validity (Barkaoui, 2010). Performance test validity is tied to raters and rubrics, in particular, because there are certain propositions that must be counted as true in order for scores to be considered valid (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014, see Standard 6.9). For example:

- “Raters attend to the criteria included in the rubrics when making their judgments (i.e., they are using appropriate criteria when they are assigning their ratings).
- Raters use the categories on the rubrics in the intended manner, applying the rubrics consistently and accurately to judge each performance (or product)” (Myford, 2012, pp. 48-49).

In this study, I focused on rater-rubric interactions, which continue to be of interest because, despite rater-training efforts, variance in rater behavior and scores persist (Lumley & McNamara, 1995; McNamara, 1996; Weigle, 2002; Weir, 2005) which may lead to reliability problems. Though the goal of rater training is to give raters a common understanding of the rubric criteria and to help raters converge



on a common understanding of scoring bands (Bejar, 2012; Roch, Woehr, Mishra, & Kieszczyńska, 2012), many studies on rater behavior have shown that raters do not always use rubrics in a consistent way (i.e., they have low intra-rater reliability). Raters do not consistently score (i.e., they have low inter-rater reliability), and they do not use the same processes to arrive at a given score (Cumming, Kantor, & Powers, 2002; Eckes, 2008; Kondo-Brown, 2002; Lumley, 2002; Orr, 2002). As Winke and Lim (2015) suggested, one potential explanation for rater behavior and problems with inter-rater reliability may be the *primacy effect*. The primacy effect is a psychological phenomenon that shows that the positionality of information in a list (e.g., a rubric) affects a listener's or reader's assignment of importance to that information (Forgas, 2011). This seems particularly relevant for helping to explain how raters pay attention to rubric criteria. Primacy may have a potential impact on inter-rater reliability on analytic rubrics. No researcher, however, has directly investigated the role of primacy in rater cognition and its potential effects on rater scoring; however, Winke and Lim (2015), posited that primacy effects were observable in their study. Thus, I investigated primacy effects in relation to rater-rubric interactions, and I examined whether they affect behavior, such as mental-rubric formation, attention to criteria, and rater scoring, when raters use an analytic rubric.

Research Questions

1. To what extent do raters show evidence of ordering effects in their mental-rubric formation after rater training?
2. To what extent do raters show evidence of ordering effects through their rubric usage during rating?
3. To what extent are raters' scores impacted by ordering effects?

Research Methodology

I employed a mixed-methods within-subjects design and included eye-tracking methodology, criteria importance surveys, criteria recall tasks, decision-making process outlines, and rater interviews. Thirty-one novice raters were randomly assigned to two groups, who, for counterbalancing purposes, were trained on two rubrics in two phases. The rubrics were a standard rubric (SR; from Polio, 2013) and a reordered rubric (RR; identical to the SR, except with categories appearing in a mirrored order to the SR). In Round 1, raters trained on one of the two rubrics and rated the same 20 essays using the rubric. The second round took place five weeks after the completion of the first. In Round 2, raters trained on the alternate rubric and re-rated the same 20 essays. Throughout the two rounds, I utilized several data-collection tools to investigate raters' cognition and behavior related to their rubric of training. Using Criteria Importance Surveys (CIS), I examined raters' beliefs about category importance. From the Criteria Recall Tasks (CRT), I examined raters' recall of the descriptors in each rubric category. With eye tracking methodology, I recorded the raters' focus on the rubric criteria during essay rating to uncover how raters used the rubric criteria based on the position of the categories. Finally, from raters' essay scores, I examined the raters' scoring consistency and severity for each rubric category.

Summary of Findings

The multiple data measures tell the same story: as novice raters train on a new rubric and assign scores using the individual categories on the rubric, the raters' behavior pertaining to the outer-most positions (e.g., left-most and right-most) seems most susceptible to ordering effects. That is, the findings of this study show that the category position affected the raters' beliefs about what criteria are the most and least important when scoring an essay, how many descriptors raters were able to recall from a category, how much attention raters paid to a category on the rubric while rating, and how



severely raters scored a given category. Additionally, the findings provided evidence that there was an interplay between the category types and category positions, resulting in either more pronounced primacy effects or leveling effects for individual rubric categories. Perhaps most importantly, there was evidence of a halo effect, in which the first category affected raters' scoring severity in the subsequent categories.

Implications

Based on the findings of this study, it would be beneficial for test designers to carefully consider the layout and ordering of analytic rubrics used in operational testing. Rubric designers could leverage ordering effects to their benefit by fronting any categories that are typically seen as less important or have lower interrater reliability scores. Test designers may also want to consider making word counts similar across categories (as done by Polio [2013] in her paper on revising the Jacobs et al. [1981] rubric) and striving for clarity and precision in each individual descriptor in order to reduce the amount of rater interpretation needed for a descriptor, as requested by Knoch (2009).

Given that raters may become more and more entrenched in their beliefs and scoring patterns when rating over long periods of time, test designers could also consider creating an online rater-training and scoring platform (see Knoch, Read, & von Randow, 2007; Wolfe, Matthews, & Vickers, 2010) which would encourage raters to pay equal attention to each rubric category. One example may be a digital platform that presents raters with a randomized, forced order of training, norming, and scoring. For each essay, the platform could randomly prompt raters to score a given category, only allowing raters to score one category at a time and input scores for the category appears on the screen. This may reduce rater's conditioning to attend most to certain categories while least to others. Additionally, many researchers advise having two raters score each essay (Elder, Barkhuizen, Knoch, & von Randow, 2007; Lumley & McNamara, 1995; Marzano, 2002; McNamara, 1996), and if raters trained and scored on categories in a random order, then pairs of raters would provide a more balanced scoring scheme and would be an additional step to mitigate any effects of primacy on scoring.

In the case that rating programs intend certain categories to be more important, those categories should be left-most, and training should indicate that the left-most categories are more important and explain why. I suspect that this is being done subconsciously in rating programs that use analytic rubrics. The rater training most likely has the new raters learn about the categories in the order they are presented (from left to right on the rubric). The rater trainers most likely work through sample scoring scenarios using the rubric from left to right, and may even unintentionally spend more time explaining the left-most categories. This ordering may have an effect on mental rubric representation, how raters view the importance of the categories, and how well certain categories are used over time. This study shows that ordering effects are real. Rater training programs now need to use that information to better design rating programs such that any ordering effects are intentional and to the betterment of the program, or the category ordering needs to be controlled so that ordering effects will not take hold and be detrimental to the rating program over time.



References

- American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US). (2014). *Standards for Educational and Psychological Testing*. Washington, D.C.: American Educational Research Association.
- Anderson, N. H. (1965). Primacy effects in personality impression formation using a generalized order effect paradigm. *Journal of Personality and Social Psychology*, 2, 1-9.
- Athey, T. R., & McIntyre, R. M. (1987). Effect of rater training on rater accuracy: Level of processing theory and social facilitation theory perspectives. *Journal of Applied Psychology*, 72, 239–244.
- Asch, S. E. (1946). Forming impressions of personality. *The Journal of Abnormal and Social Psychology*, 41(3), 258–290. <http://doi.org/10.1027/1864-9335/a000179>
- Bacha, N. (2001). Writing evaluation: What can analytic versus holistic essay scoring tell us? *System*, 29(4), 371–383.
- Baker, B. A. (2012). Individual differences in rater decision-making style: An exploratory mixed-methods study. *Language Assessment Quarterly*, 9(3), 225–248. <http://doi.org/10.1080/15434303.2011.637262>
- Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing*, 12(2), 86–107. <http://doi.org/10.1016/j.asw.2007.07.001>
- Barkaoui, K. (2008). *Effects of scoring method and rater experience on ESL essay rating processes and outcomes* (Unpublished doctoral thesis). University of Toronto, Canada
- Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, 7(1), 54–74. <http://doi.org/10.1080/15434300903464418>
- Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. *Language Testing*, 28(1), 51–75. <http://doi.org/10.1177/0265532210376379>
- Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice*, 31(3), 2–9. <http://doi.org/10.1111/j.1745-3992.2012.00238.x>
- Bond, T. G., & Fox, C. M. (2013). *Applying the Rasch model: Fundamental measurement in the human sciences*. Hove, UK: Psychology Press.
- Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment* (2nd ed.). New York, NY: McGraw-Hill.



- Brown, J. D., & Bailey, K. M. (1984). A categorical instrument for scoring second language writing skills. *Language Learning, 34*(4), 21–42.
- Brown, G. T. L., Glasswell, K., & Harland, D. (2004). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing, 9*(2), 105–121. <http://doi.org/10.1016/j.asw.2004.07.001>
- Burrows, C. (1994). Testing, testing, 1, 2, 3: An investigation of the reliability of the assessment guideline for the Certificate of Spoken and Written English. *Making Connections, 1994 ACTA-WATESOL National Conference*, 11-17.
- Carr, N. (2000). A comparison of the effects of analytic and holistic composition in the context of composition tests. *Issues in Applied Linguistics, 11*(2), 207–241.
- Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English, 18*, 65-81.
- Connor-Linton, J., & Polio, C. (2014). Comparing perspectives on L2 writing: Multiple analyses of a common corpus. *Journal of Second Language Writing, 26*, 1–9. <http://doi.org/10.1016/j.jslw.2014.09.002>
- Crano, W. D. (1977). Primacy versus recency in retention of information and opinion change. *The Journal of Social Psychology, 101*(1), 87–96.
- Crusan, D. (2015). Dance, ten; looks, three: Why rubrics matter. *Assessing Writing, 26*, 1–4. <http://doi.org/10.1016/j.asw.2015.08.002>
- Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7*(1), 31–51.
- Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating tasks: A ESL / EFL writing framework descriptive. *The Modern Language Journal, 86*(1), 67–96.
- Davis, L. (2015). The influence of training and experience on rater performance in scoring spoken language. *Language Testing, 33*(1), 117-135. <http://doi.org/10.1177/0265532215582282>
- Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*(2), 155-185. <http://doi.org/10.1177/0265532207086780>
- Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly, 9*(3), 270–292. <http://doi.org/10.1080/15434303.2011.649381>
- Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing, 24*(1), 37–64. <http://doi.org/10.1177/0265532207071511>



- Field, A. (2009). *Discovering statistics using SPSS*. Thousand Oaks, CA: Sage publications.
- Follman, J. & Anderson, J. (1967). An investigation of the reliability of five procedures for grading English themes. *Research in the Teaching of English, 1*, 190-200.
- Forgas, J. P. (2011). Can negative affect eliminate the power of first impressions? Affective influences on primacy and recency effects in impression formation. *Journal of Experimental Social Psychology, 47*(2), 425–429. <http://doi.org/10.1016/j.jesp.2010.11.005>
- Godfroid, Al., & Spino, L. A. (2015). Reconceptualizing reactivity of think-alouds and eye tracking: Absence of evidence is not evidence of absence. *Language Learning, 65*(4), 896-928.
- Goulden, N. R. (1992). Theory and vocabulary for communication assessments. *Communication Education, 41*(3), 258–269.
- Goulden, N. R. (1994). Relationship of analytic and holistic methods to rater's scores for speeches. *The Journal of Research and Development in Education, 27*, 73–82.
- Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing Second Language Writing in Academic Contexts* (pp. 241–276). Norwood, NJ: Ablex.
- Hamp-Lyons, L. (1995). Rating nonnative writing: The trouble with holistic scoring. *TESOL Quarterly, 29*, 759–762.
- Hamp-Lyons, L., & Henning, G. (1991). Communicative writing profiles: An investigation of the transferability of a multiple-trait scoring instrument across ESL writing assessment contexts. *Language Learning, 41*(3), 337–373.
- Henderson, J. M., Luke, S. G., Schmidt, J., & Richards, J. E. (2013). Co-registration of eye movements and event-related potentials in connected-text paragraph reading. *Frontiers in Systems Neuroscience, 7*, 28. <http://doi.org/10.3389/fnsys.2013.00028>
- Hendrick, C., & Costantini, A. F. (1970). Effects of varying trait inconsistency and response requirements on the primacy effect in impression formation. *Journal of Personality and Social Psychology, 15*(2), 158–164. <http://doi.org/10.1037/h0029203>
- Jacobs, H., Zingraf, S., Wormuth, D., Hartfiel, V., & Hughey, J. (1981). *Testing ESL composition: A practical approach*. Rowley, MA: Newbury House.
- Janssen, G., Meier, V., & Trace, J. (2015). Building a better rubric: Mixed methods rubric revision. *Assessing Writing, 26*(18), 1–16. <http://doi.org/10.1016/j.asw.2015.07.002>
- Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing, 26*(4), 485–505. <http://doi.org/10.1177/0265532209340186>



- Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2(2), 130–144. <http://doi.org/10.1016/j.edurev.2007.05.002>
- Kang, O. (2012). Impact of rater characteristics and prosodic features of speaker accentedness on ratings of international teaching assistants' oral performance. *Language Assessment Quarterly*, 9(3), 249–269. <http://doi.org/10.1080/15434303.2011.642631>
- Kline, R. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.
- Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*, 26(2), 275–304. <http://doi.org/10.1177/0265532208101008>
- Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing*, 16(2), 81–96. <http://doi.org/10.1016/j.asw.2011.02.003>
- Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing*, 12(1), 26–43. <http://doi.org/10.1016/j.asw.2007.04.001>
- Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, 19(1), 3–31. <http://doi.org/10.1191/0265532202lt218oa>
- Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. New York, NY: Routledge.
- Li, H., & He, L. (2015). A comparison of EFL raters' essay-rating processes across two types of rating scales. *Language Assessment Quarterly*, 12(2), 178–212. <http://doi.org/10.1080/15434303.2015.1011738>
- Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, 28(4), 543–560. <http://doi.org/10.1177/0265532211406422>
- Linacre, J. M. (1989). *Many-faceted Rasch measurement*. Chicago, IL: MESA Press.
- Luchins, A., & Luchins, E. (1970). The effects of order of presentation of information and explanatory models. *The Journal of Social Psychology*, 80, 63–70. <http://doi.org/10.1007/s13398-014-0173-7.2>
- Lumley, T. (2002). Assessment criteria in a large-scale writing test: what do they really mean to the raters? *Language Testing*, 19(3), 246–276. <http://doi.org/10.1191/0265532202lt230oa>
- Lumley T (2005) *Assessing second language writing: The rater's perspective*. New York, NY: Peter Lang.



- Lumley, T., & McNamara, T. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*(1), 54-71.
- Marzano, R. J. (2002). A comparison of selected methods of scoring classroom assessments. *Applied Measurement in Education, 15*(3), 249–268. <http://doi.org/10.1207/S15324818AME1503>
- McCray, G., & Brunfaut, T. (2016). Investigating the construct measured by banked gap-fill items: Evidence from eye-tracking. *Language Testing, 1*, 23.
- McDermott, W. L. (1986). *The scalability of degrees of foreign accent*. (Doctoral dissertation). Cornell University, Ithaca, NY.
- McNamara, T. F. (1996). *Measuring second language performance*. London, UK: Longman.
- Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility and intelligibility in the speech of second language learners. *Language Learning, 48*(1), 73–97.
- Myford, C. M. (2012). Rater cognition research: Some possible directions for the future. *Educational Measurement: Issues and Practice, 31*(3), 48–49. <http://doi.org/10.1111/j.1745-3992.2012.00243.x>
- Orr, M. (2002). The FCE speaking test: Using rater reports to help interpret test scores. *System, 30*(2), 143–154. [http://doi.org/10.1016/S0346-251X\(02\)00002-7](http://doi.org/10.1016/S0346-251X(02)00002-7)
- Polio, C. (2013). *Revising a writing rubric based on raters' comments*. Paper presented at the Midwestern Association of Language Testers (MwALT) conference, East Lansing, Michigan.
- Pollatsek, A., Reichle, E. D., & Rayner, K. (2006). Tests of the E-Z Reader model: Exploring the interface between cognition and eye-movement control. *Cognitive Psychology, 52*, 1–56.
- Rebitschek, F. G., Krems, J. F., & Jahn, G. (2015). Memory activation of multiple hypotheses in sequential diagnostic reasoning. *Journal of Cognitive Psychology, 27*(6), 780-796. <http://doi.org/10.1080/20445911.2015.1026825>
- Roch, S. G., Woehr, D. J., Mishra, V., & Kieszczyńska, U. (2012). Rater training revisited: An updated meta-analytic review of frame-of-reference training. *Journal of Occupational and Organizational Psychology, 85*(2), 370–395. <http://doi.org/10.1111/j.2044-8325.2011.02045.x>
- Sakyi, A. (2000). Validation of holistic scoring for ESL writing assessment: How raters evaluate ESL compositions. In A. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 129–152). Cambridge, UK: Cambridge University Press.
- Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing, 18*(3), 303–325. <http://doi.org/10.1191/026553201680188988>



- Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *Modern Language Journal*, 76(1), 27–33.
- Smith, D. (2000). Rater judgments in the direct assessment of competency-based second language writing ability. In G. Brindley (Ed.), *Studies in immigrant English language assessment* (pp. 159–189). Sydney, Australia: National Centre for English Language Teaching and Research, Macquarie University.
- Smith, M. (2016). *Testing the shallow structure hypothesis in L2 Japanese* (Doctoral dissertation). Michigan State University, East Lansing, Michigan.
- Solano-Flores, G., & Li, M. (2006). The use of Generalizability (G) Theory in the testing of linguistic minorities. *Educational Measurement: Issues and Practice*, 25(1), 13–22.
<http://doi.org/10.1111/j.1745-3992.2006.00048.x>
- Stuhlmann, J., Daniel, C., Dellinger, A., Kenton, R., & Powers, T. (1999). A generalizability study of the effects of training on teachers' abilities to rate children's writing using a rubric. *Reading Psychology*, 20(2), 107-127.
- Tulving, E. (2008). On the law of primacy. In M. A. Mark, J. R. Anderson, & S. M. Kosslyn (Eds.), *Memory and mind: A festschrift for Gordon H. Bower* (pp. 31–48). Hove, UK: Psychology Press.
- Tyndall, B., & Kenyon, D. M. (1996). Validation of a new holistic rating scale using Rasch multifaceted analysis. In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 39–57). Clevedon, UK: Multilingual Matters.
- Underwood, G. (1975). Perceptual distinctiveness and proactive interference in the primacy effect. *The Quarterly Journal of Experimental Psychology*, 27(2), 289–294.
<http://doi.org/10.1080/14640747508400487>
- Vaughan, C. (1991). Holistic assessment: What goes on in the raters' minds? In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111-126). Norwood, NJ: Ablex.
- Weigle, S. (1994). Effect of training on raters of ESL compositions. *Language Testing*, 11, 197–223.
- Weigle, S. (1998). Using FACETS to model rater training effects. *Language Testing*, 15(2), 263–287.
<http://doi.org/10.1191/026553298670883954>
- Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Qualitative and quantitative approaches. *Assessing Writing*, 6, 145–178.
- Weigle, S. C. (2002). *Assessing Writing*. Cambridge, UK: Cambridge University Press.
- Weigle, S. C., Boldt, H., & Valsecchi, M. I. (2003). Effects of task and rater background on the evaluation of ESL student writing. *TESOL Quarterly*, 37(2), 345–354.



- Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. New York, NY: Palgrave MacMillan.
- Winke, P., Gass, S., & Myford, C. (2012). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, 30(2), 231–252. <http://doi.org/10.1177/0265532212456968>
- Winke, P., & Lim, H. (2015). ESL essay raters' cognitive processes in applying the Jacobs et al. rubric: An eye-movement study. *Assessing Writing*, 25, 37–53. <http://doi.org/10.1016/j.asw.2015.05.002>
- Wiseman, C. (2005). *A validation study comparing an analytic scoring rubric and a holistic scoring rubric in the assessment of L2 writing samples*. (Unpublished doctoral dissertation). Teachers College, Columbia University, NY.
- Wiseman, C. S. (2012). Rater effects: Ego engagement in rater decision-making. *Assessing Writing*, 17(3), 150–173. <http://doi.org/10.1016/j.asw.2011.12.001>
- Woehr, D. J. (1994). Understanding frame of reference training: The impact of training on the recall of performance information. *Journal of Applied Psychology*, 79, 525–534.
- Wolfe, E. W., Matthews, S., & Vickers, D. (2010). The effectiveness and efficiency of distributed online, regional online, and regional face-to-face training for writing assessment raters. *The Journal of Technology, Learning and Assessment*, 10(1), 4-21.
- Yorozuya, R., & Oller, J. W. (1980). Oral proficiency scales: Construct validity and the halo effect. *Language Learning*, 30(1), 135-153.