The International Research Foundation
for English Language Education

Title of Project:
The Impact of Rater Experience and Essay Quality on Rater Behavior and Scoring

Researcher:
Ozgur Sahan
Çanakkale Onsekiz Mart University
ozgursahan66@hotmail.com

Research Supervisor:
Prof. Salim Razi
Çanakkale Onsekiz Mart University

Ozgur Sahan

Final Report

Motivation for the Research

Students' English language writing performance is assessed for a variety of purposes at Turkish universities. First, students must pass high-stakes entrance and exit exams with a writing section as part of the one-year, intensive English language preparatory program (ELPP), which students are required to complete before starting their English-medium departmental studies. Students' writing performances are also evaluated throughout their university education in order to track their progress in different courses, such as academic writing or advanced writing. Further, some universities require a good command of EFL writing as a prerequisite for exchange programs like Erasmus+ because students pursue their studies with such exchange programs in English. The task of preparing exams for the aforementioned purposes typically falls to ELPP testing units or course lecturers. However, scoring procedures do not always follow formal, predetermined steps, such as training and calibrating raters to rate the essays reliably. As such, different assessment protocols are implemented at different institutions and within the same institution. Therefore, there is a need for a standardized and sound assessment system in order to provide students with fair scorings.

Although several factors can contribute to score variations in writing performance assessments, rater-related factors can be considered central to the subjectivity of writing assessment. As one of the rater features, previous rating experience is attributed to ensuring fair judgment, placing expert scorers in a superior position throughout the evaluation processes. Yet, expertise in assessing writing does not necessarily guarantee reliable scores. Therefore, it is essential to understand the differences and commonalities in raters' reactions to essays of different qualities in order to understand better the variability of ratings. To this end, this research study focuses on two factors—scorers' rating experience and essay quality—to investigate their impact on the variability of EFL essay scores and rating behaviors in Turkish tertiary-level education. Examining assessment problems commonly seen at institutional and national levels, this research gains

177 Webster St., # 220, Monterey, CA 93940 USA
Web: www.tirfonline.org / **Email:** info@tirfonline.org

significance by investigating two main sources of variation in EFL writing assessment to establish meaningful and generalizable measurements that should be relevant beyond individual contexts.

Research Questions

The main purpose of this study was to investigate the impact of rater experience and essay quality on rater behavior and essay scores. Adopting a mixed-methods research design, the variability of ratings assigned to the essays were examined from quantitative and qualitative perspectives with the following two sets of questions.

From the quantitative aspect, the first set of questions were as follows:

1. Are there any significant differences among the analytic scores of the low- and high-quality EFL essays?
2. Are there any significant differences among the analytic scores assigned by raters with varying previous rating experience?
3. What are the sources of score variation that contribute most (relatively) to the score variability of the analytic scores of EFL essays?
4. Does the reliability (e.g., dependability coefficients for criterion-referenced score interpretations and generalizability coefficients for norm-referenced scores interpretations) of the analytic scores of raters differ based on their amount of experience?

Moreover, based on the qualitative data, the following questions were asked:

5. How do raters make decisions while rating different quality EFL essays analytically?
6. How is rating experience related to EFL raters' decision-making processes and the aspects of writing they attend to?

Research Methodology

Employing a mixed-methods research approach, this study used a convergent parallel design (Cresswell, 2011). Within this research design, the qualitative and quantitative strands of research are concurrent during the process of data collection and within the overall interpretation of results but independent of each other during the data analysis phase. Thirty-three raters voluntarily participated in this study. The participants were based in 16 different state universities in 15 different cities. All the participant raters were professionals in the field of interdisciplinary English language teaching, learning and assessment, and regular employees at the School of Foreign Languages (SFL), Foreign Languages (FL) Department, or English Language Teaching (ELT) Department at state universities in Turkey. These 33 raters were all graduates from different ELT and ELL departments in Turkey, and they have the same L1 background (Turkish). The participants varied in their professional experience in teaching and assessing EFL writing. Based on their reported rating experience, participants were divided into three groups: low-experienced (n = 13), medium-experienced (n = 10), and high-experienced raters (n = 10).

The data collection tools used in this study consisted of 50 EFL essays, a 10-point analytic scoring rubric, verbal protocols, written score explanations, and a background questionnaire. Each of these materials was designed carefully before the commencement of the main data collection phase. The essays were used to obtain both qualitative and quantitative data. Before data collection, the essays were evaluated by three independent expert raters for quality division, resulting in two sets of high-quality and low-quality texts. Raters were asked to score the essays using an analytic rubric to which they had been oriented and to provide three written score explanations for each

177 Webster St., # 220, Monterey, CA 93940 USA
Web: www.tirfonline.org / **Email:** info@tirfonline.org

essay regarding their judgments. Added to that, each rater employed think-aloud protocols (TAPs) while scoring 16 essays pre-determined by the researcher. Raters were trained on how to conduct TAPs through detailed guidelines and a sample TAP video prior to data collection. Before submitting their essay packs to the researcher, raters filled out a background questionnaire as well. A total of 9,900 scores (1,650 total scores and 8,250 sub-scores), 446 think-aloud protocols, and 5,425 written score explanations were obtained from the participants. The analysis of quantitative data relied on generalizability (G-) theory approach as well as descriptive and inferential statistics; qualitative data were analyzed through deductive and inductive coding.

Summary of Findings

Raters showed statistical differences in their scores assigned to high-quality and low-quality essays. In other words, all raters were able to distinguish low-proficient student authors from their high-proficient peers. Raters varied from one another in their ratings based on their previous rating experience. High-experienced raters and low-experienced raters displayed statistically significant differences in their total ratings of low-quality essays. Furthermore, statistically significant differences were observed between their sub-scores assigned to the mechanics component of the low-quality essays. When the scoring pattern across experience groups was examined, a positive relationship between the average scores and the amount of rater experience was observed in that more experienced raters gave higher scores to the essays than low-experienced raters did.

In addition, G-theory analysis revealed that the variance due to raters was considerably high when the ratings of high-quality and low-quality essays were evaluated separately. However, the score variability due to raters was much smaller collectively, indicating that raters showed greater differences in terms of leniency and severity within each essay quality than in the overall mixed-quality set. Fourth, an almost perfect degree of inter-rater reliability was achieved within each rater group for low-quality and mixed-quality (high- and low- quality papers together) essays, and D-studies showed that a lower number of raters would still produce scores with an acceptable level of dependability index. However, the reverse is true for high-quality essays in that low dependability coefficients were found across the three rater groups, and only if the number of raters were increased unreasonably would reliable scores be obtained for high-quality essays.

When it comes to scoring behaviors, raters displayed different decision-making strategies based on essay quality and rating experience. More experienced raters were more positive compared to less experienced raters, leading to higher essay scores respectively. Generally, raters used more interpretation strategies than judgement strategies. Raters focused more on style, grammar, and mechanics when rating low-quality essays but more on ideas, rhetoric, and their general impression of the essay when rating high-quality essays. Added to that, medium- and high-experienced raters displayed similar decision-making behaviors, while low-experienced raters differed slightly from these two more experienced groups. Low-experienced raters used more interpretation strategies than their more experienced peers whereas medium- and high-experienced raters employed judgment strategies more frequently than raters with less experience did. Medium-experienced and high-experienced raters tended to employ the same strategies while rating essays of both low- and high-quality. For both low- and high-quality papers, the low-experienced raters seemed to rely on more language-focused strategies, particularly with respect to mechanics. Across experience groups, raters displayed more language-focused strategies—such as considering punctuation, spelling, and syntax—for low-quality essays than high-quality essays.

177 Webster St., # 220, Monterey, CA 93940 USA
Web: www.tirfonline.org / **Email:** info@tirfonline.org

Implications

The findings of this study underline the need for detailed and continuous rater training even for raters with extensive rating experience. In this way, scoring gaps can be reduced between raters. Traditional rater training models can be revisited as the findings suggested that score variations between raters may be related to differentiation in certain sub-scores of writing (e.g. mechanics component), as certain raters (e.g. low-experienced raters) prioritized strategies related to such components (e.g. consider spelling and punctuation) in their think-aloud protocols and the written explanations. As such, developing a rater-training model that shifts raters' focus to all aspects covered by the scoring criteria instead of emphasizing certain traits such as grammar, content, or organization might help ensure intra- and inter-rater reliability. That is, a strategy-based rater-training model built upon the most commonly used decision-making strategies may lead raters to think similarly while evaluating EFL compositions, thus resulting in more consistent scores.

Another implication addresses double-grading protocols for institutional assessment. While many institutions use double-grading, protocols for matching the rater pairs are rarely considered. Given that high-experienced raters were found to be more lenient compared to their less-experienced peers, language programs can consider matching relatively high-experienced and relatively low-experienced raters as double-grading pairs. In other words, if two high-experienced raters are paired, they may be more likely to give higher scores to a certain writing performance, while the same essay might receive a considerably lower score if the grading is conducted by two less-experienced raters. Matching relatively high- and low-experienced raters together could compensate for these effects in double-grading situations.

Although analytic rubrics are considered more reliable and advantageous than holistic scoring, the findings showed that score variations could be observed during analytic evaluation. As such, rather than using traditional holistic and/or analytic scoring scales, developing a clear and user-friendly scale with more detailed descriptors might be helpful to reduce inconsistencies between raters. Added to that, context-bound scoring scales can be developed with specific consideration of the local, cultural, and institutional dynamics.

177 Webster St., # 220, Monterey, CA 93940 USA
Web: www.tirfonline.org / **Email:** info@tirfonline.org

References

Attali, Y. (2015). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing, 33*(1), 99-115.

Attali, Y., Lewis, W., & Steier, M. (2013). Scoring with the computer: Alternative procedures for improving the reliability of holistic essay scoring. *Language Testing, 30*(1), 125-141.

Baba, K. (2009). Aspects of lexical proficiency in writing summaries in a foreign language. *Journal of Second Language Writing, 18*, 191-208. doi:10.1016/j.jslw.2009.05.003

Bacha, N. (2001). Writing evaluation: What can analytic versus holistic essay scoring tell us? *System, 29*(3), 371-383.

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

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing, 17*(1), 1-42.

Bachman, L. F., & Palmer, A. (2010). *Language testing in practice*. Oxford, UK: Oxford University Press.

Baker, A. B. (2010). Playing with the stakes: A consideration of an aspect of the social context of a gatekeeping writing assessment. *Assessing Writing, 15*(3), 133-153.

Baker, A. B. (2012). Individual differences in rater decision-making style: An exploratory mixed-methods study. *Language Assessment Quarterly, 9*(3), 225-248.

Barkaoui, K. (2007a). Participants, texts, and processes in second language writing assessment: A narrative review of the literature. *The Canadian Modern Language Review, 64*(1), 97-132.

Barkaoui, K. (2007b). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing, 12*(2), 86-107.

Barkaoui, K. (2008). *Effects of scoring method and rater experience on ESL essay rating processes and outcomes.* Unpublished doctoral dissertation, University of Toronto, Canada.

Barkaoui, K. (2010a). Do ESL essays raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study. *TESOL Quarterly, 44*(1), 31-57.

Barkaoui, K. (2010b). Explaining ESL essay holistic scores: A multilevel modeling approach. *Language Testing, 27*(4), 515-535.

Barkaoui, K. (2010c). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly, 7*(1), 54-74.

Barkaoui, K. (2011a). Effects of marking method and rater experience on ESL scores and rater performance. *Assessment in Education: Principles, Policy & Practice, 18*(3), 279-293.

Barkaoui, K. (2011b). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. *Language Testing, 28*(1), 51-75.

Barrett, S. (2001). The impact of training on rater variability. *International Educational Journal, 2*(1), 49-58.

Barritt, L., Stock, P. L., & Clark, F. (1986). Researching practice: Evaluating assessment essays. *College Composition and Communication, 37*(3), 315-327.

Baydin, A. G. (2006). *Blank map of Republic of Turkey's provinces* [Digital image]. Retrieved from https://commons.wikimedia.org/wiki/File:BlankMapTurkeyProvinces.png

Boyatzis, R. E. (1998). *Transforming qualitative information: Thematic analysis and code development*. London, UK: SAGE.

Breland, H. M. (1983). *The direct assessment of writing skill: A measurement review (ETS Research Report No: 86-9)*. Princeton, NJ: Educational Testing Service.

Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement, 24*(4), 339-353.

Brennan, R. L. (2001a). *Generalizability theory.* New York, NY: Springer.

Brennan, R. L. (2001b). *Generalizability theory: Statistics for social science and public policy*. New York: Springer-Verlag. Retrieved from https://www.google.com.tr/search?hl=tr&tbo=p&tbm=bks&q=isbn:0387952829

Brennan, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education, 24*(1), 1-21.

Briesch, A. M., Swaminathan, H., Welsh, M., & Chafouleas, S. M. (2014). Generalizability theory: A practical guide to study design, implementation, and interpretation. *Journal of Psychology, 52*(1), 13-15.

Brown, H. D. (2004). *Language assessment: Principles and classroom practice*. New York, NY: Pearson/Longman.

Brown, J. D. (1991). Do English and ESL faculties rate writing samples differently? *TESOL Quarterly, 25*(4), 587-603.

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

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

BTU SFL (2014). *Holistic scoring scale*. Bursa Technical University School of Foreign Languages, Turkey.

Carlson, S. B., Bridgeman, B., Camp, R., & Waanders, J. (1985). *Relationship of admission test scores to writing performance of native and nonnative speakers of English* (*ETS Research Report Series GRE Board Research Report GREB No. 83-2R*). Princeton, NJ: Educational Testing Service.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing. *Research in the Teaching of English, 18*(1), 65-81.

Cohen, L., & Manion, L. (1994). *Research methods in education.* New York, NY: Routledge.

Connor-Linton, J. (1995). Looking behind the curtain: What do L2 compositions really mean? *TESOL Quarterly, 29*(4), 762-765.

Cooper, P. L. (1984). The assessment of writing ability: A review of research *(ETS Research Report Series GRE Board Research Report GREB No. 82-15R).* Princeton, NJ: Educational Testing Service.

Cresswell, J. W. (2011). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). New Delhi, India: PHI Learning Private.

Cronbach, L. J., Gleser, G. C., Nada, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York, NY: Wiley.

Cumming, A. (1989). Writing expertise and second language proficiency. *Language Learning 39*(1), 81-141.

Cumming, A. (1990). Expertise in evaluating second language composition. *Language Testing, 7*(1), 31-51.

Cumming, A., Kantor, R., & Powers, D. (2001). *Scoring TOEFL essays and TOEFL 2000 prototype tasks: An investigation into raters' decision making and development of a preliminary analytic framework (TOEFL Monograph Series, Report No: 22).* Princeton, NJ: Educational Testing Service.

Cumming, A., Kantor, R., & Powers, D. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal, 86*(1), 67-96.

Daly, J. A., & Dickson-Markman, F. (1982). Contrast effects in evaluating essays. *Journal of Educational Measurement, 19*(4), 309-316.

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, NJ: Ablex.

DeRemer, M. L. (1998). Writing assessment: Raters' elaboration of the rating task. *Assessing Writing, 5*(1), 7-29.

Diederich, P. B. (1974). *Measuring growth in English*. Urbana, IL: National Council of Teachers of English.

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative and mixed methodologies*. Oxford, UK: Oxford University Press.

Ebel, R., & Frisbie, D. A. (1991). *Essentials of educational measurement.* Englewood Cliff, NJ: Prentice Hall.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*(2), 155-185.

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly, 9*(3), 270-292.

Elbow, P. (1999). Ranking, evaluating, and liking: Sorting out three forms of judgements. In R. Straub (Ed.), *A sourcebook for responding to students writing* (pp. 175-196). Cresskill, NJ: Hampton Press.

Elorbany, R., & Huang, J. (2012). Examining the impact of rater educational background on ESL writing assessment: A generalizability theory approach. *Language and Communication Quarterly, 1*(1), 2-24.

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing, 4*(2), 139-155.

Erdosy, M. U. (2004). *Exploring variability in judging writing ability in a second language: A study for experienced raters of ESL compositions* (TOEFL Research Report RR-03-17). Princeton, NJ: Educational Testing Service.

Ferris, D. R. (1994). Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *TESOL Quarterly, 28*(2), 414-420.

Frederiksen, J. R. (1992, April). *Learning to "see:" Scoring video portfolios or "beyond the hunter-gatherer" in performance assessment*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Freedman, S. W. (1981). Influences on evaluators of expository essays: Beyond the text. *Research in the Teaching of English, 15*(3), 245-255.

Freedman, S. W. (1984). The register of student and professional expository writing: Influences on teachers' responses. In R. Beach & S. Bridwell (Eds.), *New directions in composition research* (pp. 334-347). New York, NY: Guilford Press.

Freedman, S. W., & Calfee, R. C. (1983). Holistic assessment of writing: Experimental design and cognitive and cognitive theory. In P. Mosenthal, L. Tamor, & S. A. Walmsley (Eds.), *Research on writing: Principles and methods* (pp. 75-98). New York, NY: Longman.

Fulcher, G. (2010). *Practical language testing*. London, UK: Routledge.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. London and New York, NY: Routledge.

Gao, X., & Brennan, R. L. (2001). Variability of estimated variance components and related statistics in a performance assessment. *Applied Measurement in Education, 14*(2), 191-203.

Gambetti, E., Fabbri, M., Bensi, L., & Tonetti, L. (2008). A contribution to the Italian validation of the General Decision-making Style Inventory. *Personality and Individual Differences, 44*(4), 842-852.

Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit it all? *Language Testing, 26*(4), 507-531.

Gebril, A. (2010). Bringing reading-to-write and writing-only assessment tasks together: A generalizability analysis. *Assessing Writing, 15*(2), 100-117.

Gebril, A., & Plakans, L. (2014). Assembling validity evidence for assessing academic writing: Rater reactions to integrated tasks. *Assessing Writing, 21*(2), 56-73.

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 raters' scores for speeches. *The Journal of Research and Development in Education, 27*(2), 73-82.

Greenberg, K. L. (1992). Validity and reliability issues in the direct assessment of writing. *WPA: Writing Program Administration, 16*(1-2), 7-22.

Güler, N., Uyanık, G. K., & Teker, G. T. (2012). *Genellenebilirlik kuramı*. Ankara: Pegem Akademi Yayınları.

Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 69-87). New York, NY: Cambridge University Press.

Hamp-Lyons, L. (1991). Basic concepts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 5-15). Norwood, NJ: Ablex.

Hamp-Lyons, L. (1995). Rating nonnative rating: The trouble with holistic scoring. *TESOL Quarterly, 29*(4), 759-762.

Hamp-Lyons, L. (1996). The challenges of second language writing assessment. In E. White, W. Lutz & S. Kamusikiri (Eds.), *Assessment of writing: Policies, politics, practice,* (pp. 226-240). New York, NY: Modern Language Association.

Hamp-Lyons, L., & Mathias, S. P. (1994). Examining expert judgments of task difficulty on essay tests. *Journal of Second Language Writing, 3*(1), 49-68.

Hamp-Lyons, L., & Zhang, B. W. (2001). World Englishes issues in and from academic writing assessment. In L. Flowerdew & M. Peacock (Eds.), *Research perspectives on English for academic purposes*, (pp. 101-116). Cambridge, UK: Cambridge University Press.

Han, T. (2013). *The impact of rating methods and rater training on the variability and reliability of EFL students' classroom-based writing assessments in Turkish universities: An investigation of problems and solutions.* Unpublished doctoral dissertation, Atatürk University, Turkey.

Han, T. (2017). Scores assigned by inexpert raters to different quality of EFL compositions, and the raters' decision-making behaviors. *International Journal of Progressive Education, 13*(1), 136-152.

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing* (pp. 1-28). Mahwah, NJ: Lawrence Erlbaum Associates.

Henning, G. (1991). Issues in evaluating and maintaining an ESL writing assessment program. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts*, (pp. 279-292). Norwood, NJ: Ablex.

Hinkel, E. (1994). Native and nonnative speakers' pragmatic interpretations of English texts. *TESOL Quarterly, 28*(2), 353-376.

Hinkel, E. (2003). Simplicity without elegance: Features of sentences in L1 and L2 academic texts. *TESOL Quarterly, 37*(2), 275-371.

Homburg, T. J. (1984). Holistic evaluation of ESL composition: Can it be validated objectively? *TESOL Quarterly, 18*(1), 87-108.

Huang, J. (2007). *Examining the fairness of rating ESL students' writing on large-scale assessments.* Unpublished doctoral dissertation, Queen's University, Canada.

Huang, J. (2008). How accurate are ESL students' holistic writing scores on large-scale assessments? —A generalizability theory approach. *Assessing Writing, 13*(3), 201-218.

Huang, J. (2009). Factors affecting the assessment of ESL students' writing. *International Journal of Applied Educational Studies, 5*(1), 1-17.

Huang, J. (2011). Generalizability theory as evidence of concerns about fairness in large-scale ESL writing assessments. *TESOL Journal, 2*(4), 423-443.

Huang, J. (2012). Using generalizability theory to examine the accuracy and validity of large-scale ESL writing assessment. *Assessing Writing, 17*(3), 123-139.

Huang, J., & Foote, C. J. (2010). Grading between lines: What really impacts professors' holistic evaluation of ESL graduate student writing? *Language Assessment Quarterly, 7*(3), 219 – 233.

Huang, J., & Han, T. (2013). Holistic or analytic – A dilemma for professors to score EFL essays? *Leadership and Policy Quarterly, 2*(1), 1-18.

Huang, J., Han, T., Tavano, H., & Hairston, L. (2014). Using generalizability theory to examine the impact of essay quality on rating variability and reliability of ESOL writing. In J. Huang & T. Han (Eds.), *Empirical* quantitative research in social sciences*: Examining significant differences and relationships*, (pp. 127-149). New York, NY: Untested Ideas Research Center.

Hughes, A. (2003). *Testing for language teachers*. Cambridge, UK: Cambridge University Press.

Hughes, A., & Lascaratou, C. (1982). Competing criteria for error gravity. *ELT Journal, 36*(3), 175-182.

Hughes, D. E., & Keeling, B. (1984). The use of models to reduce context effects in essay scoring. *Journal of Educational Measurement, 21*(3), 277-281.

Huot, B. A. (1990). Reliability, validity and holistic scoring: What we know and what we need to know. *College Composition and Communication, 41*(2), 201-213.

Huot, B. A. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 206-236). Cresskill, NJ: Hampton Press.

Hyland, K. (2003). *Second language writing*. New York, NY: Cambridge University Press.

177 Webster St., # 220, Monterey, CA 93940 USA
Web: www.tirfonline.org / **Email:** info@tirfonline.org

James, C. (1977). Judgements of error gravities. *ELT Journal, 31*(2), 116-124.

Jennings, M., Fox, J., Graves, B., & Shohamy, E. (1999). The test-takers' choice: An investigation of the effect of topic on language-test performance. *Language Testing, 16*(4), 426-456

Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing, 26*(4), 485-505.

Kane, M. (2008, November). *Errors of measurement, theory, and public policy.* Paper presented at the 12th Annual William H. Angoff Memorial Lecture. Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/PICANG12.pdf

Kenyon, D. (1992, February). *Introductory remarks at symposium on development and use of rating scales in language testing*. Paper presented at the 14th Language Testing Research Colloquium, Vancouver, British Columbia.

Kieffer, K. M. (1998, April). *Why generalizability theory is essential and classical test theory is often inadequate?* Paper presented at the Annual Meeting of the South Western Psychological Association, New Orleans, LA.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*(2), 275-304.

Knoch, U., Read, J., & Randow, J. V. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing, 12*(1), 26-43.

Kobayashi, H., & Rinnert, C. (1996). Factors affecting composition evaluation in an EFL context: Cultural rhetorical pattern and readers' background. *Language Learning, 46*(3), 397-437.

Kobayashi, T. (1992). Native and nonnative reactions to ESL compositions. *TESOL Quarterly, 26*(1), 81-112.

Krapels, A. R. (1990). An overview of second language writing process research. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (37-56). New York, NY: Cambridge University Press.

Kroll, B. (1990). *Second language writing: Research insights for the classroom*. New York, NY: Cambridge University Press.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics 33*(1), 159-174.

Leckie, G., & Baird, J. A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement, 48*(4), 399-418.

Lee, Y.-W., Kantor, R., & Mollaun, P. (2002, April). *Score dependability of the writing and speaking sections of new TOEFL.* Paper presented at the annual meeting of National Council on Measurement in Education, New Orleans, LA. Abstract retrieved from ERIC. (*ERIC No. ED464962)*

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.

Linn, R. L., & Burton, E. (1994). Performance-based assessments: Implications of task specificity. *Educational Measurement: Issues and Practice, 13*(1), 5-8.

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.

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. New York, NY: Peter Lang.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*(1), 54-71.

Mackey, A., & Gass, S. (2005). *Second language research: Methodology and design*. Mahwah, NJ: Lawrence Erlbaum.

McColly, W. (1970). What does educational research say about the judging of writing ability? *The Journal of Educational Research, 64*(4), 148-156.

McNamara, T. F. (1996). *Measuring second language performance.* London and New York, NY: Addison Wesley Longman.

McNamara, T. F. (2000). *Language testing.* Oxford, UK: Oxford University Press.

Milanovic, M., Seville, N., & Shuhong, S. (1996). A study of the decision-making behavior of composition markers. In M. Milanovic & N. Seville (Eds.), *Performance testing, cognition, and assessment: Selected papers from the 15th Language Testing Colloquium (LTRC), Cambridge and Arnhem* (pp. 92-114). Cambridge, UK: Cambridge University Press.

Myers, M. (1980). *A procedure for writing assessment and holistic scoring*. National Council of Teachers of English, Urbana, IL.

Najimy, N. C. (1981). *Measure for measure: A guidebook for evaluating students' expository writing.* Urbana, IL: National Council of Teachers of English.

Perkins, K. (1983). On the use of composition scoring techniques, objective measures, and objective tests to evaluate ESL writing ability. *TESOL Quarterly, 17*(4), 651-671.

Plakans, L. (2007). *Second language writing and reading-to-write assessment tasks: A process study*. Unpublished doctoral dissertation. The University of Iowa.

Pula, J. J., & Huot, B. A. (1993). A model of background influences on holistic raters. In M. Williamson & B. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations*, (pp. 237-265). Gresskill, NJ: Hampton Press.

Quality Manual (2015, 15 September). *Bursa Technical University School of Foreign Languages Quality Manual.* Retrieved from http://depo.btu.edu.tr/dosyalar/ydyo/Dosyalar/SFL%20-%20Quality%20Manual%20-%2023%2011%202017%284%29.pdf

Raimes, A. (1990). The TOEFL test of written English: Causes for concern. *TESOL Quarterly, 24*(3), 427-442.

Reid, J., & O'Brien, M. (1981, March). *The application of holistic grading in an ESL writing program.* Paper presented at the Annual Convention of Teachers of English to Speakers of Other Languages. Detroit, MI. (ERIC Document Reproduction Service No. ED 221 044).

Reid, J. M. (1993). *Teaching ESL writing*. Englewood Cliffs, NJ: Prentice-Hall.

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing, 15*(1), 18-39.

Rinnert, C., & Kobayashi, H. (2001). Differing perceptions of EFL writing among readers in Japan. *The Modern Language Journal, 85*(2), 189-209.

Russikoff, K. A. (1995, March). *A comparison of writing criteria: Any differences?* Paper presented at the Annual Meeting of the Teachers of English to Speakers of Other languages, Long Beach, CA.

Saeidi, M., & Rashvand Semiyari, S. (2011). The impact of rating methods and task types on EFL learners' writing scores. *Journal of English Studies, 1*(4), 59-68.

Sakyi, A. A. (2000). Validation of holistic scoring for ESL writing assessment: How raters evaluate ESL compositions. A. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 129-152). Cambridge, UK: Cambridge University Press.

Sakyi, A. A. (2003). *A study of the holistic scoring behaviors of experienced and novice ESL instructors.* Unpublished doctoral dissertation, University of Toronto, Canada.

Santos, T. (1988). Professors' reactions to the academic writing of nonnative-speaking students. *TESOL Quarterly, 22*(1), 69-90.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing, 22*(1), 1-30.

Schoonen, R., Vergeer, M., & Eiting, M. (1997). The assessment of writing ability: Expert readers versus lay readers. *Language Testing, 14*(2), 157-184.

Scott, S. G., & Bruce, R. A. (1995). Decision-making style: The development and assessment of a new measure. *Educational and Psychological Measurement, 55*(5), 818-831.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of educational Measurement, 30*(3), 215-232.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A premier.* Newbury Park, CA: Sage.

Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing, 18*(3), 303-325.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal, 76*(1), 27-33.

Song, B., & Caruso, I. (1996). Do English and ESL faculty differ in evaluating the essays of native English-speaking, and ESL students? *Journal of Second Language Writing, 5*(2), 163-182.

Spicer, D. P., & Sadler-Smith, E. (2005). An examination of the general decision making style questionnaire in two UK samples. *Journal of Managerial Psychology, 20*(2), 137-149.

Stalnaker, J. M., & Stalnaker, R. C. (1934). Reliable reading of essay tests. *The School Review, 42*(8), 599-605.

Suen, H. (1990). *Principles of test theories.* Hillsdale, NJ: Lawrence Erlbaum.

Sweedler-Brown, C. O. (1985). The influence of training and experience on holistic essay evaluation. *English Journal, 74*(5), 49-55.

Şahan, Ö. (2016a, June 20). *Rubric orientation session* [Video file]. Retrieved from https://www.youtube.com/watch?v=AKPnsdt4nuo

Şahan, Ö. (2016b, June 23). *A sample think-aloud protocol* [Video file]. Retrieved from https://www.youtube.com/watch?v=hoJxNZFdT4Q

Şahan. Ö., & Razı, S. (2017, June). *The impact of rater experiences and essay quality on rater behavior and rating scores*. Paper presented at the 16th Symposium on Second Language Writing, Assessing Second Language Writing, Bangkok, Thailand.

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal, 49*(1), 3-12.

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. 111-125). Norwood, NJ: Ablex.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*(2), 197-223.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263-87.

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing, 6*(2), 145-178.

Weigle, S. C. (2002). *Assessing writing.* Cambridge, UK: Cambridge University Press.

Weigle, S. C. (2004). Integrating reading and writing in a competency test for non-native speakers of English. *Assessing Writing, 9*(1), 27-55.

Weir, C. J. (2005). *Language testing and validation.* Hampshire, UK: Palgrave McMillan.

White, E. M. (1994). *Teaching and assessing writing: Recent advances in understanding, evaluating, and improving student performance*. San Francisco, CA: Jossey-Bass Publishers.

Wolfe, E. F., & Feltovich, B. (1994, April). *Learning to rate essays: A study of scorer cognition.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Wolfe, E. W. (2005). Uncovering rater's cognitive processing and focus using think-aloud *protocols. Journal of Writing Assessment, 2*(1), 37-56.

Wolfe, E. W., Kao, C., & Ranney, M. (1998). Cognitive differences in proficient and non-proficient essay scorers. *Written Communication, 15*(4), 465-492.

Yang, Y. (2001). *Chinese interference in English writing: Cultural and linguistic differences.* (ERIC Document Reproduction Service No. ED 461 992).