**Title of Project:**
Rater Expertise in a Second Language Speaking Assessment: The Influence of Training and Experience

**Researcher:**
Larry Davis
University of Hawaiʻi, Mānoa
larrydavis98@yahoo.com

**Research Supervisor:**
Dr. John Norris

Larry Davis

---

**Project Summary:**

In tests of speaking it is the rater who translates a performance into a score; it is therefore the rater who is the final arbiter of what a given score means. Accordingly, to understand the meaning of a score we must understand how and why raters make particular scoring decisions. While considerable effort has been devoted to the analysis of raters' scoring patterns, many questions remain regarding how people learn to make judgments of complex phenomena such as language ability. For example, it has been reported that as essay raters gain experience they tend to be more internally consistent but do not necessarily become more consistent with other raters (Lim, 2009; Weigle, 1994, 1998). However, it is less clear how decision making changes with training or experience to actually produce such increased intra-rater reliability. Understanding how raters develop the expertise to reliably and accurately score speaking responses has obvious implications for the design of rater training and scoring procedures.

Moreover, studies in the area of rater decision making tend to be descriptive in nature, and within speaking assessment little has been done to investigate the basic mechanisms behind raters' decision making. A common assumption seems to be that scoring judgments are made by applying a scoring rubric to a speaking performance, which nicely fits the widely cited definition of measurement as the application of a rule to an observation (Stevens, 1946). However, a different view has recently emerged from the field of psychophysics and behavioral economics which asserts that magnitude judgments are made by comparing the observation at hand with other examples in the environment or memory (Laming, 2004; Stewart, Chater, & Brown, 2006). Within this framework judgments are seen as fundamentally relative in nature rather than being the application of an absolute rule.

The overall purpose of this dissertation was to extend our understanding of rater expertise within a speaking test context. The primary focus included an investigation of the ways in which rater scoring patterns and behavior change with training and experience, as well as the characteristics of scoring behavior and cognition associated with raters showing more or less desirable patterns in scoring. In addition, an initial effort was made to examine the possibility that judgments of language ability are relative in nature. The following research questions were examined:

1. What effects do training and experience have on scoring patterns and scoring behavior of raters?
2. What features of scoring behavior and cognition distinguish more-proficient and less-proficient raters?
3. Can rater decision making be understood in terms of a relative view of judgment?

To examine these questions, 20 experienced teachers of English scored recorded examinee responses from the TOEFL iBT speaking test prior to training and in three sessions following training, over a period of about three weeks. In the first session previously scored example responses were the only materials given to raters; raters awarded scores on the basis of their native scoring criteria along with whatever information they could take from the examples. This was followed by training which included exposure to the scoring rubric, review of example responses with scoring explanations, and practice scoring with feedback given after each score (the feedback being a previously established reference score for the same response). For the final three scoring sessions raters were simply asked to review the scoring rubric and listen to a set of example responses at the start of each session. Raters scored 100 responses in each session along with an additional 20 responses where they verbally reported what they were thinking as they listened to an examinee response and made a scoring decision.

For Research Question 1, scores from each rater were analyzed for consistency in scoring patterns and agreement with other raters, as well as for accuracy (agreement with reference scores). Various aspects of raters' interaction with the scoring materials were also recorded to determine if certain behaviors, such as the time taken to reach a scoring decision, were associated with the reliability and accuracy of scores. Somewhat surprisingly, rater severity and internal consistency (measured via Rasch analysis) were already of a standard typical for operational language performance tests in the first scoring session, before raters had received training or even seen the scoring rubric. This result suggests that the example responses provided in the first session may have been adequate for raters to apply the rating scale, at least for the experienced teachers used as raters in the study. Training played a role as well, however, resulting in increased agreement between raters and improved agreement with established reference scores. Additional experience gained after training appeared to have little effect on rater scoring patterns, although agreement with reference scores continued to increase somewhat.



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

For Research Question 2, less-, intermediate-, and more-proficient raters were identified on the basis of scoring consistency and accuracy and compared in terms of behavior and thought processes while scoring. More proficient raters reviewed example responses more often and took longer to make scoring decisions, suggesting the possibility that rater behavior while scoring may influence the accuracy and reliability of scores. Analyses of rater's verbal reports found no obvious differences in the frequency with which raters mentioned various language features while scoring, and considerable individual variation was seen. However, following training, mention of language features related to topic development increased for all groups, probably in response to specific instructions for evaluating topical development provided in the training. The frequency of comments on the scoring process also increased following training, and overall the training appeared to influence both the attention raters gave to certain language features and the scoring process more generally.

For Research Question 3 the extent to which scoring decisions are made by comparing examinee responses was investigated. First, scores produced by each rater were examined for sequence effect, or the tendency for a score to be correlated with the score given to the previous test taker. In the first study, small but significant correlations were found between scores produced in succession, supporting the view that scoring judgments are relative in nature. Second, an experiment was conducted where raters were presented with pairs of responses and asked to decide if the responses deserved the same or different scores; the prediction was that it would be easier to distinguish between similar responses when these could be compared side-by-side. Contrary to this prediction, discrimination between examinees was actually poorer than when the same responses were individually scored. Many raters also felt that the comparison task was cognitively demanding because details of the first response had to be held in mind while judging the second response. This memory load may have been a confounding factor and complicates interpretation of the experiment results, but from a practical standpoint underscores the challenge of judging speaking ability through direct comparison of responses.

The study has a number of implications for assessment practice. First, the results highlight the importance of examples for aligning rater perceptions to the rating scale. With only the exemplars for guidance, raters in the study were able to achieve reasonably good score reliability and accuracy in the first scoring session. So, it may be useful to make examples available while scoring and encourage raters to refer to the examples often. The results also raise the question of the degree to which descriptions of performance found in scoring rubrics actually influence raters' scores. While exposure to the scoring rubric appeared to influence rater perceptions and improve accuracy in scoring, it seems likely that scoring decisions also rely on the examples used to develop an understanding of the rating scale. Accordingly, scoring rubrics may represent



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

only a partial description of the scoring criteria actually employed in a language performance test.

The results also demonstrate that training has value, as indicated by the improved scoring performance seen after raters completed a series of typical training activities. Subsequent experience resulted in little further improvement in scoring performance, suggesting that unguided experience may not necessarily be very effective in developing scoring expertise, at least once a rater has reached a certain level of performance. This view is consistent with the broader literature on expertise, which suggests that elite levels of performance can be obtained only through focused practice, and simple repetition is not enough.

Finally, in this study I evaluated a possible mechanism for how scoring decisions are made; namely, that judgments of language ability are relative in nature. While the results of the current study are inconclusive, several findings suggest that further examination of this theoretical position is warranted, including the finding that (a) scoring judgments were influenced by similar examples of performance available in the immediate environment, as indicated by a sequence effect; and (b) greater use of exemplars for comparison was associated with scoring accuracy/consistency in more-proficient raters. Although the empirical findings produced so far are quite modest, the current study introduces to the domain of language assessment a new, and testable, framework for thinking about the way raters make judgments of language ability.



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

## References

Adobe Systems. (2006). Adobe LiveCycle Designer (Version 8.0) [Computer Software]. San Jose, CA: Adobe Systems.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika 43*, 561–573.

Alderson, J. C. (1991). Bands and scores. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s* (pp. 71–86). London, England: Modern English Publications and The British Council.

Attali, Y. (2011). Sequential effects in essay ratings. *Educational and Psychological Measurement, 71*, 68–79.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, England: Cambridge University Press.

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing, 12*, 238–257.

Barkaoui, K. (2007). Participants, texts, and processes in ESL/EFL essay tests: A narrative review of the literature. *Canadian Modern Language Review, 64*, 99–134.

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

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

Barnwell, D. (1989). 'Naive' native speakers and judgements of oral proficiency in Spanish. *Language Testing, 6*, 152–163.

Black, B., & Bramley, T. (2008). Investigating a judgemental rank-ordering method for maintaining standards in UK examinations. *Research Papers in Education, 23*, 357–373.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing, 20*, 89–110.

Bridgeman, B. (2012). Why bother with research when we have common sense? *R&D Connections, 20*, 1–8. Retrieved from: www.ets.org/Media/Research/pdf/RD_Connections_20.pdf

Bridgeman, B., Powers, D., Stone, E., & Mollaun, P. (2011). TOEFL iBT speaking test scores as

indicators of oral communicative language proficiency. *Language Testing, 29*, 91–108.

Brooks, V. (2012). Marking as judgment. *Research Papers in Education, 27*, 63–80.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing, 12*, 1–15.

Brown, A. (2007). An investigation of the rating process in the IELTS oral interview. In L. Taylor & P. Falvey (Eds.), *IELTS collected papers* (pp. 98–139). Cambridge, England: Cambridge University Press.

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test–taker performance on English-for-academic-purposes speaking tasks* (TOEFL Monograph No. 29, RR-05-05). Retrieved from: https://www.ets.org/Media/Research/pdf/RR-05-05.pdf

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

Brown, J. D. (2001). *Using surveys in language programs*. Cambridge, England: Cambridge University Press.

Brown, J. D., & Ahn, R. C. (2011). Variables that affect the dependability of L2 pragmatics tests. *Journal of Pragmatics, 43*, 198–217.

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

Camerer, C. F., & Johnson, E. J. (1991). The process-performance paradox in expert judgment: How can experts know so much and predict so badly? In K. A. Ericsson and J. Smith (Eds.), *Toward a general theory of expertise: Prospects and limits* (pp. 195–217). Cambridge, England: Cambridge University Press.

Carey, M. D., & Mannell, R. H. (2009). The contribution of interlanguage phonology accommodation to inter-examiner variation in the rating of pronunciation in oral proficiency interviews. *IELTS Research Reports, 9*, 217–236.

Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing, 12*, 54–71.

Cho, D. (1999). A study on ESL writing assessment: Intra–rater reliability of ESL compositions. *Melbourne Papers in Language Testing, 8*(1), 1–24.

Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement, 37*, 163–178.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*, 213–220.

Connor, U. M., & Carrell, P. L. (1993). The interpretation of tasks by writers and readers in holistically rated direct assessment of writing. In J. G. Carson & I. Leki (Eds.), *Reading in the composition classroom* (pp. 141–160). Boston, MA: Heinle & Heinle.

Connor-Linton, J. (1995). Crosscultural comparison of writing standards: American ESL and Japanese EFL. *World Englishes, 14*, 99–115.

Cowon Systems. (2011). JetAudio Player (Version 8.0.16.2000) [Computer software]. Seoul, Korea: Cowon Systems.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7*, 31–51.

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

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

Delaruelle, S. (1997). Text type and rater decision-making in the writing module. In G. Brindley & G. Wigglesworth (Eds.), *Access: Issues in language test design and delivery* (pp. 215–242). Sydney, Australia: National Centre for English Language Teaching and Research, Macquarie University.

Diederich, P. B., French, J. W., & Carlton, S. T. (1961). Factors in judgements of writing ability. *Research Bulletin* (ETS Research Bulletin No. RB-61-15). Princeton, NJ: Educational Testing Service (ERIC Document Reproduction Service ED 002172).

Douglas, D., & Myers, R. (2000). Assessing the communication skills of veterinary students: Whose criteria? In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (pp. 60–81). Cambridge, England: Cambridge University Press.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly, 2*, 197–221.

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

Eckes, T. (2009). On common ground? How raters perceive scoring criteria in oral proficiency testing. In A. Brown & K. Hill (Eds.), *Tasks and criteria in performance assessment: Proceedings of the 28th Language Testing Research Colloquium* (pp. 43–73). Frankfurt, Germany: Peter Lang.

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

Educational Testing Service. (2007). *The official guide to the new TOEFL iBT* (2nd ed.). New

York, NY: McGraw Hill.

Educational Testing Service. (2008). *TOEFL iBT public use dataset* [data files, score prompts, scoring rubrics]. Princeton, NJ: Educational Testing Service.

Educational Testing Service. (2011). *TOEFL iBT sample questions* [data and software]. Retrieved from http://www.ets.org/toefl/ibt/prepare/sample_questions

Elder, C. (1993). How do subject specialists construe classroom language proficiency? *Language Testing, 10*, 235–254.

Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly, 2*, 175–196.

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*, 37–64.

Englehard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*, 93–112.

Enright, M. K., Bridgeman, B., Eignor, D., Lee, Y. W., & Powers, D. E. (2008). Prototyping measures of listening, reading, speaking, and writing. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 145–186). New York, NY: Routledge.

Erdosy, M. U. (2004). Exploring variability in judging writing ability in a second language: A study of four experienced raters of ESL compositions (TOEFL Research Report No. 70, RR-03-17). Retrieved from: http://www.ets.org/Media/Research/pdf/RR-03-17.pdf

Ericsson, K. A. (2006). An introduction to The Cambridge Handbook of Expertise and Expert Performance: Its development, organization, and content. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge Handbook of Expertise and Expert Performance* (pp. 3–19). Cambridge, England: Cambridge University Press.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (2nd ed.). Cambridge, MA: MIT Press.

Fahim, M., & Bijani, H. (2011). The effects of rater training on raters' severity and bias in second language writing assessment. *Iranian Journal of Language Testing, 1*, 1–16.

Feltovich, P. J., Prietula, M. J., & Ericsson, K. A. (2006). Studies of expertise from psychological perspectives. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge Handbook of Expertise and Expert Performance* (pp. 41–67). Cambridge, England: Cambridge University Press.

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice, 40*, 532–538.

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

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed.). Hoboken, NJ: John Wiley & Sons.

Freedman, S. W., & Calfee, R. C. (1983). Holistic assessment of writing: Experimental design 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. (2003). *Testing second language speaking*. Harlow, UK: Longman.

Furneaux, C., & Rignall, M. (2007). The effect of standardization–training on rater judgements for the IELTS writing module. In L. Taylor & P. Falvey (Eds.), *IELTS Collected Papers: Research in speaking and writing assessment* (pp. 422–445). Cambridge, England: Cambridge University Press.

Galloway, V. B. (1980). Perceptions of the communicative efforts of American students of Spanish. *Modern Language Journal, 64*, 428–433.

Gass, S. M., & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah, NJ: Lawrence Erlbaum Associates.

Glaser, R., & Chi, M. T. H. (1988). Overview. In M. T. H. Chi, R. Glaser, & M. J. Farr (Eds.), *The nature of expertise* (pp. xv–xxviii). Hillsdale, NJ: Lawrence Erlbaum Associates.

Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook*. Cambridge, England: Cambridge University Press.

Gui, M. (2012). Exploring differences between Chinese and American EFL teachers' evaluations of speech performance. *Language Assessment Quarterly, 9*, 186–203.

Gwet, K. L. (2011). AgreeStat (Version 2011.2) [Computer software]. Gaithersburg, MD: Advanced Analytics.

Hales, L. W., & Tokar, E. (1975). The effect of the quality of preceding responses on the grades assigned to subsequent responses to an essay question. *Journal of Educational Measurement, 12*, 115–117.

Harding, L. & Ryan, K. (2009). Decision making in marking open-ended listening test items: The case of the OET. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 7*, 99–114.

Hatch, E., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguists*. Boston, MA: Heinle & Heinle.

Hill, K. (1996). Who should be the judge? The use of non–native speakers as raters on a test of English as an international language. *Melbourne Papers in Language Testing, 5*(2), 29–49.



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

Hsieh, C. N. (2011). Rater effects in ITA testing: ESL teachers' versus American undergraduates' judgments of accentedness, comprehensibility, and oral proficiency. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 9*, 47–74.

Hughes, D. C., & Keeling, B. (1984). The use of model essays to reduce context effects in essay scoring. *Journal of Educational Measurement, 21*, 277–281.

Hughes, D. C., Keeling, B., & Tuck, B. F. (1980). The influence of context position and scoring method on essay scoring. *Journal of Educational Measurement, 17*, 131–135.

Huot, B. (1988). *The validity of holistic scoring: A comparison of the talk–aloud protocols of expert and novice holistic raters*. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses Database. (UMI No. 8817872)

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60*, 237–263.

Huot, B. (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.

Jesteadt, W., Luce, R. D., & Green, D. M. (1977). Sequential effects in judgments of loudness. *Journal of Experimental Psychology: Human Perception and Performance, 3*, 92–104.

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

Kang, O. (2008). Ratings of L2 oral performance in English: Relative impact of rater characteristics and acoustic measures of accentedness. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 6*, 181–205.

Kim, H. J. (2011). *Investigating raters' development of rating ability on a second language speaking assessment.* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 3448033)

Kim, Y. H. (2009a). A G-theory analysis of rater effect in ESL speaking assessment. *Applied Linguistics, 30*, 435–440.

Kim, Y. H. (2009b). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing, 26*, 187–217.

Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior – a longitudinal study. *Language Testing, 28*, 179–200.

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



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

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

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

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing, 19*, 3–31.

Laming, D. (1984). The relativity of 'absolute' judgements. *British Journal of Mathematical and Statistical Psychology, 37*, 152–183.

Laming, D. (1997). *The measurement of sensation*. Oxford, England: Oxford University Press.

Laming, D. (2003). Marking university examinations: Some lessons from psychophysics. *Psychology Learning and Teaching, 3*, 89–96.

Laming, D. (2004). *Human judgment: The eye of the beholder*. London, England: Thomson.

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

Lantolf, J. P., & Frawley, W. (1985). Oral-proficiency testing: A critical analysis. *Modern Language Journal, 69*, 337–345.

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

Lee, Y. W., & Kantor, R. (2005). *Dependability of ESL writing test scores: Evaluating prototype tasks and alternative rating schemes*. (TOEFL Monograph MS-31). Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/RR-05-14.pdf

Lee, Y. W. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing, 23*, 131–166.

Lim, G. S. (2009). *Prompt and rater effects in second language writing performance assessment.* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses Database. (UMI No. 3392954)

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*, 543–560.

Linacre, J. M. (2007a). *A user's guide to FACETS*. Chicago, IL: Publisher.

Linacre, J. M. (2007b). FACETS Rasch measurement computer program (Version 3.62.0)

[Computer software]. Chicago, IL: Publisher.

Linacre, J. M. (2010). *A user's guide to FACETS/MINIFAC Rasch-model computer programs*. Chicago, IL: Publisher.

Lumley, T. (1998). Perceptions of language-trained raters and occupational experts in a test of occupational English language proficiency. *English for Specific Purposes, 17,* 347–367.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*, 246–276.

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

Lumley, T., & Brown, A. (2005). Research methods in language testing. In E. Hinkle (Ed.), *Handbook of research in second language teaching and learning* (pp. 833–856). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Lunz, M. E., & Stahl, J. A. (1990). Judge consistency and severity across grading periods. *Evaluation & the Health Professions, 13*, 425–444.

Luoma, S. (2004). *Assessing speaking*. Cambridge, England: Cambridge University Press.

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15*, 158–180.

Macmillan, N. A., & Creelman, C. D. (2005). Detection theory: A user's guide (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research, 50*, 940–967.

Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.

Matthews, W. J., & Stewart, N. (2009*).* Psychophysics and the judgment of price: Judging complex objects on a non-physical dimension elicits sequential effects like those in perceptual tasks. *Judgment and Decision Making, 4*, 64–81.

May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing, 26*, 397–421.

McIntyre, P. N. (1993). *The importance and effectiveness of moderation training on the reliability of teacher assessments of ESL writing sample.* (Unpublished master's thesis).

University of Melbourne, Melbourne, Australia.

McNamara, T.F. (1996). *Measuring second language performance*. London, England: Longman.

McNamara, T. F., & Lumley, T. (1997). The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational settings. *Language Testing, 14*, 140–156.

Meadows, M., & Billington, L. (2005). *A review of the literature on marking reliability* (Report produced for the National Assessment Agency). Retrieved from http://collections-r.europarchive.org/tna/20081118111836/http://naa.org.uk/libraryAssets/media/Review_of_the_literature_on_marking_reliability.pdf

Meiron, B. E. (1998). *Rating oral proficiency tests: A triangulated study of rater thought processes.* (Unpublished master's thesis). California State University Los Angeles, Los Angeles, California.

Mendelsohn, D., & Cumming, A. (1987). Professors' ratings of language use and rhetorical organizations in ESL compositions. *TESL Canada Journal, 5*, 9–26.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp.13-103). New York, NY: American Council on Education and Macmillan.

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

Mislevy, R. J. (2010). Some implications of expertise research for educational assessment. *Research Papers in Education, 25*, 253–270.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher, 23*, 5–12.

Myford, C. M., & Wolfe, E. W. (2000). *Monitoring sources of variability within the Test of Spoken English assessment system*. (TOEFL Research Report No. 65, RR-00-6). Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/RR-00-06-Myford.pdf

Myford, C. M. & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*, 386–422.

Myford, C. M., Marr, D. B., & Linacre, J. M. (1996). *Reader Calibration and Its Potential Role in Equating for the Test of Written English.* (TOEFL Research Report No. 52, RR-95-40). Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/RR-95-40.pdf

Nádas, R., & Suto, I. (2010). Speed isn't everything: A study of examination marking. *Educational Studies, 36*, 115–118.

Neter, J., Wasserman, W., & Kutner, M. H. (1985). *Applied linear statistical models: Regression, analysis of variance, and experimental designs*. Homewood, IL: Richard D. Irwin.

Norris, J. M. (1996). *A validation study of the ACTFL guidelines and the German speaking test.* (Unpublished master's thesis). University of Hawai'i, Honolulu, Hawaii.

Norris, J. M., Brown, J. D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments.* Honolulu, HI: University of Hawai'i, National Foreign Language Resource Center.

O'Loughlin, K. (1992). Do English and ESL teachers rate essays differently? *Melbourne Papers in Language Testing, 1*(2), 19–44.

Orr, M. (2002). The FCE Speaking test: Using rater reports to help interpret test scores. *System, 30*, 143–154.

O'Sullivan, B., & Rignall, M. (2007). Assessing the value of bias analysis feedback to raters for the IELTS writing module. In L. Taylor & P. Falvey (Eds.), *IELTS Collected Papers: Research in speaking and writing assessment* (pp. 446–478). Cambridge, England: Cambridge University Press.

Ozer, D. J. (1993). Classical psychophysics and the assessment of agreement and accuracy in judgments. *Journal of Personality, 64*, 739–767.

Papajohn, D. (2002). Concept mapping for rater training. *TESOL Quarterly, 36*, 219–233.

Phillips, J. K., Klein, G., & Sieck, W. R. (2004). Expertise in judgment and decision making: A case for training intuitive decision skills. In D. J. Koehler & N. Harvey (Eds.), *Blackwell Handbook of Judgment and Decision Making* (pp. 297–315). Hoboken, NJ: Wiley-Blackwell.

Plous, S. (1993). *The psychology of judgment and decision making*. New York, NY: McGraw-Hill.

Pollack, I. (1952). The information of elementary auditory displays. *Journal of the Acoustical Society of America, 24*, 745–749.

Pollitt, A. (2004, June). *Let's stop marking exams*. Paper presented at the IAEA Conference, Philadelphia. Retrieved from http://www.cambridgeassessment.org.uk/ca/digitalAssets/113942_Let_s_Stop_Marking_Exams.pdf

Pollitt, A., & Crisp, V. (2004, September). *Could comparative judgements of script quality replace traditional marking and improve the validity of exam questions?* Paper presented

at the British Educational Research Association Annual Conference, Manchester, England. Retrieved from http://cambridgeassessment.org.uk/ca/digitalAssets/113798_Could_comparative_judgements.pdf

Pollitt, A., & Murray, N. L. (1996). What raters really pay attention to. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium (LTRC), Cambridge and Arnhem* (pp. 74–91). Cambridge, England: Cambridge University Press.

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

Quellmalz, E. (1980). *Problems in stabilizing the judgment process* (CSE Report No. 136). University of California, Los Angeles, National Center for Research on Evaluation, Standards, & Student Testing. Retrieved from http://www.cse.ucla.edu/products/reports/R136.pdf

*Random House Webster's College Dictionary* (2nd ed.). (1997). New York, NY: Random House.

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

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*, 413–428.

Sakyi, A. A. (2000). Validation of holistic scoring for ESL writing assessment: How raters evaluate compositions. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (pp. 129–152). Cambridge, England: Cambridge University Press.

Sakyi, A. A. (2003). *The study of the holistic scoring behaviours of experienced and novice ESL instructors*. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses Database. (UMI No. NQ78033)

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

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*, 465–493.

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

Shanteau, J. (1992). The psychology of experts: An alternative view. In G. Wright & F. Bolger



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

(Eds.), *Expertise and decision support* (pp. 11–23). New York, NY: Plenum.

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

Shaw, S. (2002). The effect of training and standardization on rater judgement and inter-rater reliability. *Research Notes, 9,* 13–17. Retrieved from http://www.cambridgeesol.org/rs_notes/rs_nts8.pdf

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

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*, 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. 159–190). Sydney, Australia: National Centre for English Language Teaching and Research, Macquarie University.

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*, 163–182.

Spear, M. (1997). The influence of contrast effects upon teachers' marks. *Educational Research, 39*, 229–233.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 677–680.

Stewart, N., Brown, G. D. A., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological Review, 112*, 881–911.

Stewart, N., Chater, N., & Brown, G. D. A. (2006). Decision by sampling. *Cognitive Psychology, 53*, 1–26.

Stricker, L. J. (1997). *Using just noticeable differences to interpret Test of Spoken English scores* (TOEFL Research Report 58, RR 97–4). Princeton, NJ: Educational Testing Service.

Suto, W. M. I., & Greatorex, J. (2008). What goes through an examiner's mind? Using verbal protocols to gain insights into the GCSE marking process. *British Educational Research Journal, 34*, 213–233.

Tsui, A. B. M. (2003).*Understanding expertise in teaching: Case studies of ESL teachers*. Cambridge, England: Cambridge University Press.

Turner, C. E., & Upshur, J. A. (2002). Rating scales derived from student samples: Effects of the scale maker and the student sample on scale content and student scores. *TESOL*

*Quarterly, 36*, 49–70.

Upshur, J. A., & Turner, C. E., (1999). Systematic effects in the rating of second-language speaking ability: Test method and learner discourse. *Language Testing, 16*, 82–111.

Vann, R. J., Meyer, D. E., & Lorenz, F. O. (1984). Error gravity: A study of faculty opinion of ESL errors. *TESOL Quarterly, 18*, 427–440.

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111–125). Norwood, NJ: Ablex.

Ward, L. M. (1979). Stimulus information and sequential dependencies in magnitude estimation and crossmodality matching. *Journal of Experimental Psychology: Human Perception and Performance, 5*, 444–459.

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

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*, 263–287.

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

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

Wier, C. C., Jesteadt, W., and Green, D. M. (1977). Frequency determination as a function of frequency and sensation level. *Journal of the Acoustical Society of America, 61*, 178–184.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing, 10*, 305–319.

Winke, P., Gass, S., & Myford, C. (2011). The relationship between raters' prior language study and the evaluation of foreign language speech samples (TOEFL iBT Research Report No. 16, RR-11-30). Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/RR-11-30.pdf

Wolfe, E. W. (1995). *A study of expertise in essay scoring.* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses Database. (UMI No. 9621424)

Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing, 4*, 83–106.

Wolfe, E. W. (2006). Uncovering rater's cognitive processing and focus using think-aloud protocols. *Journal of Writing Assessment, 2*, 37–56.

Wolfe, E. W., Kao, C. W., & Ranney, M. (1998). Cognitive differences in proficient and

nonproficient essay scorers. *Written Communication, 15*, 465–492.

Xi, X., & Mollaun, P. (2009). *How do raters from India perform in scoring the TOEFL iBT™ speaking section and what kind of training helps?* (TOEFL iBT Research Report No. 11, RR-09-31). Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/RR-09-31.pdf

Xi, X. (2007). Evaluating analytic scoring for the TOEFL Academic Speaking Test (TAST) for operational use. *Language Testing, 24*, 251–286.

Xiph.org. (2006). *Speex: A free codec for free speech* [Webpage]. Retrieved from http://www.speex.org/

Yamanaka, H. (2005). Using generalizability theory in the evaluation of L2 writing. *JALT Journal, 27*, 169–185.

Yoshida, H. (2004). *An analytic instrument for assessing EFL pronunciation*. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses Database. (UMI No. 9315947)

Zhang, Y., & Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing, 28*, 31-50.