Title of Project:
Rater effects in ITA testing: ESL teachers' versus American undergraduates' judgments of oral proficiency, accentedness, and comprehensibility

Researcher:
Ching-Ni Hsieh
Michigan State University
chsieh@ets.org

Research Supervisor:
Dr. Paula Winke

Ching-Ni Hsieh

Summary of Research Findings

This dissertation project contributes to the field of language assessment—the TIRF's research priority for which the project was funded—in three main aspects. First of all, this study examines the quality of human ratings in oral performance assessment within a high-stakes testing context. Such an investigation aligns with TIRF's goal to ensure that English as a second/foreign language is tested in a manner that is demonstrably fair and effective. Secondly, this study includes linguistically naïve undergraduate raters, who are the main stakeholders in the international teaching assistants (ITAs) testing situation. Results of the study help identify comprehension difficulties that undergraduate students experience with ITAs and shed light on ITA testing. Thirdly, this study employs a mixed-method design in data collection and analysis, a research design TIRF encourages. This methodology helps provide a deeper understanding of rater effects on performance assessment as different aspects of rater behaviors are elicited by different methods.

In this study, I compared ratings of oral proficiency, accentedness, and comprehensibility awarded by English-as-a-second-language (ESL) teachers and American undergraduates to understand the role of rater background in measures of ITA speech. I also explored factors to which raters attended while they evaluated ITA oral proficiency.

The research questions guiding this study included:

1. Do ESL teachers and American undergraduate students differ in the severity with which they evaluate potential ITAs' oral proficiency, accentedness, and comprehensibility, respectively, and if so, to what extent?

2. What factors draw raters' attention while the raters evaluate potential ITAs' oral proficiency? Are different factors more or less salient to different rater groups?

To answer the research questions, I had 13 ESL teachers and 32 U. S. undergraduate students evaluated 28 ITA candidates' oral responses to the Speaking Proficiency English Assessment Kit (SPEAK), whose results are used to screen ITAs at Michigan State University. Raters assigned ratings online and provided written comments regarding the factors they took into consideration while judging candidates' oral proficiency. After raters completed their ratings, one-on-one follow-up interviews were conducted to probe raters' reasons for making

rating decisions. The scores assigned by the two groups of raters were subject to different statistical analyses to determine whether the two groups of raters differed in severity. Written comments and interview protocols were analyzed to understand raters' rating orientations.

In response to the first research question, the results suggest that the rater groups did not differ in the severity they exercised when they evaluated the examinees' oral proficiency. This finding is backed by the overall results of the multiple quantitative analyses, including the descriptive statistics of the raw scores, the classification of ITAs assignments, the FACETS analyses, and the Mann-Whitney *U* tests. However, there were significant, between-group differences between the two groups' ratings on accentedness and comprehensibility. The undergraduate raters were more severe when they judged the examinees' foreign accents. They also perceived a significantly higher level of difficulty in comprehending the examinees' speech.

The second research question delved into *why* raters with different backgrounds may differentially rate the speech of ITAs. By coding the written comments, I identified six main rating categories the raters reported they employed: linguistic resources, phonology, fluency, content, global assessment, and other, nonlinguistic factors. Raters' attention to the first four rating categories was broken down further. For example, within the linguistic resources category, raters made comments on the examinees' use of grammar, vocabulary, expression, and textualization. Within the phonology category, the examinees' pronunciation, intonation, rhythm and stress, and foreign accent were all sources of attention. As far as fluency is concerned, raters judged the responses based on the repetitions or self-repair patterns and the speech rate of the speakers. In terms of content, raters noted whether the examinees fulfilled the task requirements, the ideas that the examinees produced, and the organization of the responses. Nonlinguistic factors included test-taking strategies, voice quality, and examinees' emotions.

The quantitative comparisons of the written comments and the qualitative analysis of the interview protocols further helped determine the extent to which rater groups differed in the rating criteria they utilized. The results of these separate analyses converged, indicating that the ESL teachers and the undergraduates attended to several aspects of the linguistic dimensions in the examinees' speech differently. Specifically, the results suggest that the teacher raters commented more frequently on a variety of linguistic features than did the undergraduates. The undergraduates, on the other hand, appeared to evaluate the examinees' oral performances more impressionistically. The interview data reveal that many undergraduates were not familiar with the rating criteria for judging the SPEAK examinees and, thus, they made their rating decisions solely through their appraisal of whether they felt a particular examinee was qualified to be an ITA, or whether they would like the speaker to be their TA—a criterion not on the rating rubric. In either case, the data appear to suggest that undergraduate raters consider their personal feelings, perhaps even their fears, and their possible future experiences as students in ITA classes in judging ITA speech. They may tend to err on the side of caution and be more severe on accent and comprehensibility, regardless of oral proficiency, in anticipation of possibly having the test taker as a teacher in the future.

One pedagogical implication of this study has to do with the training of undergraduate students with regard to how to listen to accented speech. Many undergraduates have a general tendency to feel anxious about listening to foreign-accented speech due to their limited experience interacting with L2 speakers, or their lack of confidence in their own abilities to communicate or understand foreigners. However, it is not fair to say that ITAs are always at fault when there is a communication breakdown in the classroom. ITA programs should consider not only offering pronunciation instructions to ITAs but also make available training workshops that

teach undergraduates how to listen to and process accented L2 speech. These workshops can help reduce undergraduates' anxiety while they listen to or converse with L2 speakers or their ITAs. Even through very limited training, undergraduate students can increase their ability to comprehend accented speech and enhance their willingness to talk with L2 speakers.

This study also has implications for ITA testing. Undergraduates are important stakeholders in the ITA testing context and should be included as a part of the ITA screening process. As the results show, although the ratings of oral proficiency assigned by the undergraduates were comparable to those assigned by the ESL teachers, significant differences on the ratings of accentedness and comprehensibility across the rater groups were found in this study. The undergraduate raters were more severe in terms of accent and comprehensibility judgments. It can be argued that the ESL teachers' judgments of the examinees' oral performances in terms of accent and comprehensibility were more lenient than the undergraduates' because the ESL teachers paid more attention to specific, linguistic features in the speech samples, while the undergraduates tended to base their ratings more on accent (and for the undergraduates, the heavier the accent, the worse the comprehensibility of the speech) and overall *feel*. A main finding of this study is that undergraduates may not be able be act as impartial judges, even with extensive training, because they have something at stake—the possibility to be taught by ITAs who they cannot understand. This study's results suggest that ITA programs should *avoid* having undergraduates as official raters, but rather use them to check the threshold of what undergraduates may consider as incomprehensible speech. On the other hand, ITA testing program should not underestimate undergraduates' abilities to adapt and comprehend ITAs whose speech falls within that "grey" zone (between what undergraduate raters would call incomprehensible, but what expert ESL teachers would call comprehensible, since research has shown that through very limited training, undergraduate students can increase their ability to comprehend accented speech and willingness to talk with L2 speakers (e.g., Derwing & Munro, 2009; Derwing et al., 2002). Therefore, ITA testing programs should not fear such a gap. Nevertheless, the potential difference in severity between the official ITA testing raters and the undergraduates should still be constantly monitored, carefully evaluated, researched, and controlled.

References

Anderson-Hsieh, J., & Koehler, K. (1988). The effect of foreign accent and speaking rate on native speaker comprehension. *Language Learning, 38*(4), 561-613.

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

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing, 12*(2), 238-257.

Bailey, K. M. (1983). Foreign teaching assistants at U.S. universities: Problems in interaction and communication. *TESOL Quarterly, 17*(2), 308-310.

Bailey, K. M. (1984a). A typology of teaching assistants. In K. M. Bailey, F. Pialorsi & J. Zukowski/Faust (Eds.), *Foreign teaching assistants in U.S. universities* (pp. 110-125). Washington D.C.: National Association for Foreign Student Affairs.

Bailey, K. M. (1984b). The "foreign TA problem". In K. M. Bailey, F. Pialorsi & J. Zukowski/Faust (Eds.), *Foreign teaching assistants in US universities* (pp. 3-16). Washington, DC: National Association for Foreign Student Affairs.

Bannon, P. (2005, June 24). Brilliant instructors, imperfect English. *The New York Times*. Retrieved from http://www.nytimes.com

Barkaoui, K. (2010a). Do ESL essay 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*, 54-74.

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

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

Bauer, G. (1996). Addressing special considerations when working with international teaching assistants. In J. D. Nyquist & D. H. Wulff (Eds.), *Working effectively with graduate assistants* (pp. 85-103). London: Sage Publications.

Bejar, I. I. (1985). *A preliminary study of raters for the Test of Spoken English (TOEFL*

Research Report RR-85-5). Princeton, NJ: Educational Testing Service. Retrived from http://www.ets.org/Media/Research/pdf/RR-85-05.pdf

Binghadeed, N. (2008). Acoustic analysis of pitch range in the production of native and nonnative speakers of English. *The Asian EFL Journal Quarterly, 10*, 96-113.

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

Bresnahan, M. J., Ohashi, R., Nebashi, R., Liu, W. Y., & Shearman, S. M. (2002). Attitudinal and affective response toward accented English. *Language and Communication, 22*, 171-185.

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

Brown, A. (2007). An investigation of the rating process in the IELTS oral interview. In M. Milanovic & C. J. Weir (Eds.), *IELTS collected papers: Research in speaking and writing assessment* (pp. 98-141). Cambridge: Cambridge University Press.

Brown, A., & Hill, K. (2007). Interviewer style and candidate performance in the IELTS oral interview. In L. Taylor & P. Falvey (Eds.), *IELTS collected papers: Research in speaking and writing assessment* (pp. 37-61). Cambridge: Cambridge University Press.

Brown, A., Iwashita, N., & McNamara, T. F. (2005). *An examination of rater orientations and test-taker performance on English-for-academic-purpose speaking tasks* (TOEFL Research Report RR-05-05). Princeton, NJ: Educational Testing Service. Retrived from http://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, K. (1992). American college student attitudes toward non-native instructors. *Multilingua, 11*(3), 249-265.

Brown, K., Fishman, P., & Jones, N. (1990). *Legal and policy issues in the language proficiency assessment of international teaching assistants*. Houston: University of Houston Law Center.

Bryd, P., & Constantinides, J. C. (1992). The language of teaching mathematics: Implications for training ITAs. *TESOL Quarterly, 26*(1), 163-167.

Cargile, A. C., & Giles, H. (1997). Understanding language attitudes: Exploring listener affect and identity. *Language and Communication, 17*(3), 195-217.

Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing, 12*, 16-35.

Chalhoub-Deville, M., & Wigglesworth, G. (2005). Rater judgment and English language speaking proficiency. *World Englishes, 24*(3), 383-391.

Chiang, S.-Y. (2009). Dealing with communication problems in the instructional interactions between international teaching assistants and American college students. *Language and Education, 23*(5), 461-478.

Clarke, J., & Swinton, C. (1980). *The Test of Spoken English as a measure of communicative ability in English-medium instructional settings* (TOEFL Research Report RR 80-33). Princeton, NJ: Educational Testing Service. Retrived http://www.ets.org/Media/Research/pdf/RR-80-33.pdf

Congdon, P., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement, 37*(2), 163-178.

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

Council of Graduate Schools (2007). 2006 CGS international graduate admissions survey. Retrieved from http://www.cgsnet.org/portals/0/pdf/R_Intlenrl06_III.pdf

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 ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal, 86*(1), 67-96.

Dalle, T. S., & Inglis, M. J. (1989, March). *What really affects undergraduates' evaluations of nonnative teaching assistant's teaching?* Paper presented at the meeting of Teachers of English to Speakers of Other Languages, San Antonio, TX.

Davies, C., Tyler, A., & Koran, J. (1989). Face-to-face with native speakers: An advanced training class for international teaching assisstants. *English for Specific Purposes, 8*, 139-153.

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 Center for English Language Teaching and Research, Macquarie University.

Derwing, T. M. (1990). Speech rate is no simple matter: Rate adjustment and NS-NNS communicative success. *Studies in Second Language Acquisition, 12*, 303-313.

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition, 20*, 1-16.

Derwing, T. M., & Munro, M. J. (2001). What speaking rates do non-native listeners prefer? *Applied Linguistics, 22*(3), 324-337.

Derwing, T. M., & Munro, M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly, 39*(3), 379-398.

Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching, 42*(4), 476-490.

Derwing, T. M., Munro, M. J., & Wiebe, G. (1998). Evidence in favor of a broad framework for pronunciation instruction. *Language Learning, 48*, 383-410.

Derwing, T. M., Rossiter, M. J., & Munro, M. J. (2002). Teaching native speakers to listen to foreign-accented speech. *Journal of Multilingual and Multicultural Development, 23*(4), 245-259.

Dick, R. C., & Robinson, B. M. (1994). Oral English proficiency requirements for ITAs in U.S. colleges and universities: An issue in speech communication. *JACA, 2*(1), 77-86.

Douglas, D. (1994). Quantity and quality in speaking test performance. *Language Testing, 11*(2), 125-144.

Douglas, D., & Smith, J. (1997). *Theoretical underpinnings of the Test of Spoken English revision project* (TOEFL Monograph Series RM-97-2). Princeton, NJ:Educational Testing Service. Retried from http://www.ets.org/Media/Research/pdf/RM-97-02.pdf

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. (in press). Many-facet Rasch measurement. In T. S. (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*. Strasbourg, France: Council of Europe/Language Policy Division.

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

Engelhard, G. J., & Myford, C. M. (2003). *Monitoring faculty consultant performance in the advnaced placement English literature and composition program with a many-faceted Rasch model* (College Board Research Report No. 2003-1). Princeton, NJ: Educational

Testing Service. Retried from http://www.ets.org/Media/Research/pdf/RR-03-01-Engelhard.pdf

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

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). Thousand Oaks, CA: Sage.

Finder, A. (2005, June 25). Unclear on American campus: What the foreign teacher said. *The New York Times*. Retried from http://www.nytimes.com

Flege, J. E. (1988a). Factors affecting degree of perceived foreign accent in English sentences. *Journal of the Acoustical Society of America, 84*(1), 70-79.

Flege, J. E. (1988b). The production and perception of foreign language speech sounds. In H. Winitz (Ed.), *Human communication and its disorders: A review-1988* (pp. 224-401). Norwood, NJ: Ablex.

Flege, J. E., & Fletcher, K. L. (1992). Talker and listener effects on degree of perceived foreign accent. *Journal of the Acoustical Society of America, 91*, 370-389.

Flege, J. E., Munro, M. J., & MacKay, I. R. A. (1995). Factors affecting strength of perceived foreign accent in a second language. *Journal of the Acoustical Society of America, 106*, 2973-2987.

Fox, W. S., & Gay, G. (1994). Functions and effects of international teaching assistants. *Review of Higher Education, 18*, 1-24.

Gass, S. M., & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning, 34*, 65-89.

Gorsuch, G. J. (2003, December). The educational cultures of international teaching assistants and U.S. universities. *TESOL-EJ, 7*(3). Retrieved from http://www.tesl-ej.org/wordpress/

Gravois, J. (2005, April 8). Teach impediment. *The Chronicle of Higher Education.* Retried from http://chronicle.com/section/Home/5

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

Hadden, B. L. (1991). Teacher and nonteacher perceptions of second-language communication. *Language Learning, 41*(1), 1-24.

Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly, 38*(2), 201-223.

Hinofotis, F. B., & Bailey, C. M. (1981). American undergraduates' reactions to the communication skills of foreign teaching assistants. In J. C. Fisher, M. A. Clarke & J. Schachter (Eds.), *On TESOL '80: Building bridges: Research and practice in teaching English as a second language* (pp. 120-133). Washington, DC: TESOL.

Hoekje, B., & Linnell, K. (1994). "Authenticity" in language testing: Evaluating spoken language tests for international teaching assistants. *TESOL Quarterly, 28*(1), 103-126.

Hoekje, B., & Williams, J. (1992). Communicative competence and the dilemma of international teaching assistant education. *TESOL Quarterly, 26*(2), 243-269.

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods, 5*(1), 64-86.

Isaacs, T. (2008). Towards defining a valid assessment criterion of pronunciation proficiency in non-native English-speaking graduate students. *The Canadian Modern Language Review, 64*(4), 555-580.

Iwashita, N., Brown, A., McNamara, T. F., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29*, 24-49.

Jia, C. L., & Bergerson, A. A. (2008). Understanding the international teaching assistant training program: A case study at a northwestern research university. *International Education, 37*(2), 77-129.

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

Juffs, A. (1990). Tone, syllable structure and interlanguage phonology: Chinese learners' stress errors. *International Review of Applied Linguistics, 28*, 99-117.

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.

Kang, O. (2010). Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness. *System, 38*, 301-315.

Kim, Y.-H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing, 26*(2), 187-217.

Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (pp. 1-14). Cambridge: Cambridge University Press.

Kunnan, A. J. (2005). Towards a model of test evaluation: Using the Test Fairness and the Test Context Frameworks. In L. Taylor & C. J. Weir (Eds.), *Multilingualism and assessment* (pp. 229-251). Cambridge: Cambridge University Press.

Landa, M. (1988). Training international students as teaching assistants. In J. A. Mestenhauser & G. Marty (Eds.), *Culture, learning, and the disciplines: Theory and practice in cross-cultural orientation* (pp. 50-57). Washington, DC: National Association for Foreign Student Affairs.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.

Linacre, J. M. (1998). Rating, judges and fairness. *Rasch Measurement Transactions, 12*, 630-631.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J. M. (2010). FACETS (Version 3.67) [Computer Software]. Chicago: WINSTEPS.com.

Lindemann, S. (2002). Listening with an attitude: A model of native-speaker comprehension of non-native speakers in the United States. *Language in Society, 31*, 419-441.

Lippi-Green, R. (1997). *English with an accent: Language, ideology, and discrimination in the United States*. New York: Routledge.

LoCastro, V., & Tapper, G. (2008). International teaching assistants and teacher identity. *Journal of Applied Linguistics, 3*(2), 185-218.

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: Lang.

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 and the Health Professions, 13*, 425-444.

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*(2), 158-180.

McCracken, G. (1988). *The long interview*. London: Sage.

McNamara, T. F. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing, 7*(1), 52-75.

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

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

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

Meiron, B. E. (1998, April). Rating oral proficiency tests: A triangulated study of rater thought processes. Paper presented at the meeting of Language Testing Research Colloquium, Monterey, CA.

Meiron, B. E., & Schick, L. S. (2000). Ratings, raters and test performance: An exploratory study. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (pp. 153-176). Cambridge: Cambridge University Press.

Mendelsohn, D., & Cumming, A. (1987). Professor's ratings of language use and rhetorical organizations in ESL compositions. *TESL Canada Journal, 5*(1), 9-26.

Merrylees, B., & McDowell, C. (2007). A survey of examiner attitudes and behavior in the IELTS oral interview. In L. Taylor & P. Falvey (Eds.), *IELTS collected papers: Research in speaking and writing assessment* (pp. 142-184). Cambridge: Cambridge University Press.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: ACE/Macmillan.

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

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*. Thousand Oaks: Sage.

Monoson, P. K., & Thomas, C. F. (1993). Oral English proficiency policies for faculty in U.S. higher education. *Review of Higher Education, 16*, 127-140.

Morley, J. (1994). A multidimensional curriculum design for speech-pronunciation instruction. In J. Morley (Ed.), *Pronunciation pedagogy and theory* (pp. 64-91). Alexandra, VA: Teachers of English to Speakers of Other Languages.

Munro, M. J. (2008). Foreign accent and speech intelligibility. In J. G. H. Edwards & M. L. Zampini (Eds.), *Phonology and second language acquisition* (pp. 193-218). Philadelphia, PA: John Benjamins.

Munro, M. J., & Derwing, T. M. (1994). Evaluations of foreign accent in extemporaneous and read material. *Language Testing, 11*(3), 253-266.

Munro, M. J., & Derwing, T. M. (1995a). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 41*(1), 73-97.

Munro, M. J., & Derwing, T. M. (1995b). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech, 38*, 289-306.

Munro, M. J., & Derwing, T. M. (1998). The effects of speaking rate on listener evaluations of native and foreign-accented speech. *Language Learning, 48*, 159-182.

Munro, M. J., & Derwing, T. M. (1999). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 49*, 285-310.

Munro, M. J., & Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech: The role of speaking rate. *Studies in Second Language Acquisition, 23*, 451-468.

Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition, 28*, 111-131.

Munro, M. J., Derwing, T. M., & Sato, K. (2006). Salient accents, covert attitudes: Consciousness-raising for pre-service second language learners. *Prospect, 21*(1), 67-79.

Muthuswamy, N., Smith, R., & Strom, R. B. (2004, May). "Understanding the problem": International teaching assistants and communication. Paper presented at the meeting of International Communication Association, New Orleans, LA.

Myford, C. M., & Wolfe, E. W. (2000). *Monitoring sources of variability within the Test of Spoken English assessment system* (TOEFL Research Report RR-00-06). Princeton, NJ: Educational Testing Service. Retried 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-faceted Rasch measurement: Part I. *Journal of Applied Measurement, 4*, 386-422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-faceted Rasch measurement: Part II. *Journal of Applied Measurement, 5*, 189-227.

Nakatsuhara, F. (2008). Inter-interviewer variation in oral interview tests. *ELT Journal, 62*, 266-275.

O'Loughlin, K. (1994). The assessment of writing by English and ESL teachers. *Australian Review of Applied Linguistics, 17*, 23-44.

O'Loughlin, K. (2007). An investigation into the role of gender in the IELTS oral interview. In L. Taylor & P. Falvey (Eds.), *IELTS collected papers: Research in speaking and writing assessment* (pp. 63-95). Cambridge: Cambridge University Press.

Okoth, E., & Mupinga, D. M. (2007, February). An evaluation of the international graduate teaching assistants training program. Paper presented at the meeting of The Academy of Human Resource Development International Research Conference in the Americas, Indianapolis, IN. Retried from http://www.eric.ed.gov/PDFS/ED504334.pdf

Oppenheim, N. (1996, April). Undergraduates learning from nonnative English-speaking teaching assistants. Paper presented at the meeting of the American Educational Research Association, New York. Retried from http://www.eric.ed.gov/PDFS/ED504334.pdf

Oppenheim, N. (1997, March). How international teaching assistant programs can prevent lawsuits. Paper presented at the meeting of the American Educational Research Association, Chicago, IL. Retried from http://www.eric.ed.gov/PDFS/ED408886.pdf

Oppenheim, N. (1998, March). Undergraduates' assessment of international teaching assistants' communicative competence. Paper presented at the meeting of the Teachers of English to Speakers of Other Languages. Retried from http://www.eric.ed.gov/PDFS/ED423783.pdf

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

Orth, J. L. (1983). *University undergraduate evaluational reactions to the speech of foreign teaching assistants.* Unpublished doctoral dissertation, University of Texas at Austin, Austin, Texas.

Papajohn, D. (2002). Concept mapping for rater training. *TESOL Quarterly, 36*(2), 219-233.

Patton, M. Q. (1990). *Qualitative evaluation and research methods*. Newbury Park: Sage.

Pica, T., Barnes, G. A., & Finger, A. G. (1990a). *Discourse and performance of international teaching assistants*. New York, NY: Newbury House.

Pica, T., Barnes, G. A., & Finger, A. G. (1990b). *Teaching matters: Skills and strategies for international teaching assistants*. New York: Newbury House.

Pickering, L. (2001). The role of tone choice in improving ITA communication in the classroom. *TESOL Quarterly, 35*(2), 233-255.

Pickering, L. (2004). The structure and function of intonational paragraphs in native and

nonnative speaker instructional discourse. *English for Specific Purposes, 23*(1), 19-43.

Piske, T., MacKay, I. R. A., & Flege, J. E. (2001). Factors affecting degree of foreign accent in an L2: A review. *Journal of Phonetics, 29*, 191-215.

Plakans, B. S. (1997). Undergraduates' experiences with and attitudes toward international teaching assistants. *TESOL Quarterly, 31*(1), 95-118.

Plough, I. C., Briggs, S. L., & Van Bonn, S. (2010). A multi-method analysis of evaluation criteria used to assess the speaking proficiency of graduate student instructors. *Language Testing, 27*(2), 235-260.

Powers, D. E., Shedl, M. A., Wilson-Leung, S., & Butler, F. A. (1999). *Validating the revised TSE® against a criterion of communicative success* (TOEFL Research Report 99-05). Princeton: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/RR-99-05.pdf

Rao, N. (1995, May). The Oh No! Syndrom: A Language Expectation Model of undergraduates' negative reactions toward foreign teaching assistants. Paper presented at the meeting of the International Communication Association, Albuquerque, NM. Retried from http://www.eric.ed.gov/PDFS/ED384921.pdf

Reed, D. J., & Cohen, A. D. (2001). Revisiting raters and ratings in oral language assessment. In C. Elder (Ed.), *Experimenting with uncertainty: Essays in honour of Alan Davies* (pp. 82-96). Cambridge: Cambridge University Press.

Rounds, P. L. (1987). Characterizing successful classroom discourse for NNS teaching assistant training. *TESOL Quarterly, 21*(4), 643-672.

Rubin, D. (1992). Nonlanguage factors affecting undergraduates' judgments of nonnative English-speaking teaching assistants. *Research in Higher Education, 33*(4), 511-531.

Rubin, D., & Smith, K. A. (1990). Effects of accent, ethnicity, and lecture topic on undergraduates' perceptions of nonnative English-speaking teaching assistants. *International Journal of Intercultural Relations, 14*, 337-353.

Ruderman, A. (2000, December 27). Colleges are moving to ensure English fluency in teaching assistants. *The New York Times*. Retried from http://www.nytimes.com

Saif, S. (2002). A needs-based approach to the evaluation of the spoken language ability of international teaching assistants. *The Canadian Journal of Applied Linguistics, 5*, 145-167.

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, UK: Cambridge University Press.

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

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

Schmid, P. M., & Yeni-Komshian, G. H. (1999). The effects of speaker accent and target predictability on perception of mispronunciations. *Journal of Speech, Language, and Hearing Research, 42*, 56-64.

Sebastian, R., & Ryan, E. B. (1985). Speech cues and social evaluation: Markers of ethnicity, social class, and age. In H. Giles & R. N. St. Clair (Eds.), *Recent advances in language, communication and social psychology* (pp. 112-143). London, England: Lawrence Erlbaum.

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

Teicher, S. A. (2005, April 18). When you can't understand the teacher. *The Christian Science Monitor*. Retrieved from http://www.csmonitor.com/2005/0418/p11s02-ussc.html

Trofimovich, P., & Baker, W. (2006). Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition, 28*(1), 1-30.

Tyler, A. (1992). Discourse structure and the perception of incoherence in international teaching assistants' spoken discourse. *TESOL Quarterly, 26*(4), 713-729.

Tyler, A., Jefferies, A., & Davies, C. E. (1988). The effect of discourse structuring devices on listener perceptions of coherence in non-native university teachers' spoken discourse. *World Englishes, 7*, 101-110.

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*(2), 263-287.

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

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke, UK: Palgrave Macmillan.

Williams, J. (1992). Planning, discourse marking, and the comprehensibility of international teaching assistants. *TESOL Quarterly, 26*(4), 693-711.

Winke, P., Gass, S. M., & Myford, C. M. (2011). *The relationship between raters' prior language study and the evaluation of foreign language speech samples*. Princeton: Educational Testing Service.

Xi, X. (2008). *Investigating the criterion-related validity of the TOEFL Speaking Scores for ITA screening and setting standards for ITAs* (TOEFL iBT Research Report 08-02). Princeton: Educational Testing Service. Retried from http://www.ets.org/Media/Research/pdf/RR-08-02.pdf

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 09-31). Princeton, NJ: Educational Testing Service. Retried from http://www.ets.org/Media/Research/pdf/RR-09-31.pdf

Yook, E. L., & Albert, R. D. (1999). Perceptions of international teaching assistants: The interrelatedness of intercultural training, cognition, and emotion. *Communication Education, 48*, 1-17.

Yule, G., & Hofffman, P. (1990). Predicting success for international teaching assistants in a U.S. university. *TESOL Quarterly, 24*(2), 227-243.

Zhao, Y. (1997). The effects of listener' control of speech rate on second language comprehension. *Applied Linguistics, 18*, 49-68.

Zielinski, B. (2006). The intelligibility cocktail: An interaction between speaker and listener ingredients. *Prospect, 21*(1), 22-45.

Zielinski, B. (2008). The listener: No longer the silent partner in reduced intelligibility. *System, 36*, 69-84.