LRN Research Proposal 2013

Evaluating Inter-Rater Reliability in Speaking Assessments: Handling Sparse Data under Generalizability Theory Framework

Chih-Kai Lin

Department of Educational Psychology

University of Illinois at Urbana-Champaign, USA

April 10, 2013

Introduction

Performance-based language assessment serves as an alternative to traditional multiple-choice item formats. It offers a more direct measure of a person's language proficiency in speaking. The advent of performance-based language testing is motivated by validity concerns regarding the extent to which assessment tasks resemble real-life tasks (Bachman & Palmer, 1996; Chapelle, Enright, & Jamieson, 2008). In addition, valid inferences about examinee ability are contingent upon score reliability from multiple raters. Hence, methods in evaluating inter-rater reliability are needed to determine the utility of any speaking assessments. Multiple sources of systematic variation (i.e., facets) can affect score reliability (Fulcher, 2003). To this end, generalizability theory or G theory (Brennan, 2001) is a powerful analytical tool that provides information about how much variation is explicable by different facets (e.g., raters and tasks) and how score reliability changes if we alter rating designs (e.g., increasing the number of raters).

Inter-Rater Reliability under Generalizability Theory Framework

G theory has been widely used in speaking assessments to investigate score reliability with respect to variation attributed to different facets in a rated test (e.g., Brown & Ahn, 2011; Lee, 2006; Lynch & McNamara, 1998). Despite its popularity, ideal applications of G theory require fully-crossed measurement designs; that is, each examinee response is scored by all raters. Such designs may not be feasible in many operational settings because it may be more cost-effective to assign different batches of responses to groups of raters. As a result, unbalanced designs are common in practice.

In cases of unbalanced designs or sparse datasets, two common methods based on Analysis of Variance (ANOVA) have been applied under the G-theory framework to estimate variance components. These estimated variance components are then used to compute score reliability. First, *raters* are treated as a random facet (e.g., Xi, 2007); henceforth as the *rater* method. Second, *ratings* are treated as a random facet (e.g., Bachman, Lynch, & Mason, 1995); henceforth as the *rating* method. The two methods differ not only in the specifications of the random facets but also in the estimation procedures of variance components. The *rater* method identifies blocks of fully-crossed sub-datasets and estimates the variance components based on a weighted average across the sub-datasets (Chiu & Wolfe, 2001). The *rating* method forces an unbalanced design to be a fully-crossed one by conceptualizing individual *ratings*, irrespective of

which raters, as a random facet. The variance components are then estimated via the usual ANOVA procedures for any fully-crossed designs.

Research Questions

Clearly, the *rating* method is computationally less complex but achieves its simplicity at the expense of rater information by assuming that scoring variability is similar across all raters, which may not always be the case when a mixture of novice and experienced raters participate in scoring. On the other hand, the *rater* method retains rater information by giving different weights to groups of raters; nonetheless, it requires higher computational sophistication. Given the two methods, the fundamental issue here rests on whether the methodological approaches can achieve precise estimation of variance components as these estimates are the building blocks in investigating inter-rater reliability under G theory.

The proposed study aims to address the following two research questions about the *rater* and *rating* methods:

1. Does one method yields more precise reliability estimates than the other? If so, under what condition(s) is one method preferred over the other?
2. Based on results of research question 1, how many trained raters are required to achieve acceptable reliability in the speaking component of an LRN examination?

Research Design and Method

The proposed study consists of a simulation study and an empirical study. The goal of the simulation study is to investigate estimation precision. The empirical study builds on results of the simulation study and offers practical recommendations for the speaking component of an LRN assessment product. Operational data from any LRN examinations can be used in this study so long as the spoken responses are double-scored by trained raters.

Simulation Study

A Monte Carlo simulation will be conducted. Data will be simulated based on a one-facet random effect model: $X_{pr} = \mu + \alpha_p + \beta_r + \varepsilon_{pr,e}$. For example, the score (X_{pr}) of person *p*, judged by rater *r*, is the sum of the overall mean (μ) and the three random-effect components associated with persons, raters and errors. The three random-effect components will be generated independently from three normal distributions, respectively, where $\alpha_p \sim N(0, \sigma_p^2)$, $\beta_r \sim N(0, \sigma_r^2)$, and $\varepsilon_{pr,e} \sim N(0, \sigma_e^2)$. True parameters for the variance components (i.e., σ_p^2, σ_r^2, and σ_e^2) will be

selected based on previous empirical research on speaking assessments (e.g., Bachman et al., 1995; Lee, 2006; Lynch & McNamara, 1998; Xi, 2007).

Three levels of sample sizes: 50, 100 and 200, three levels of sparseness: 50%, 75% and 87.5%, and three scenarios of rater variability will be chosen in the simulation study; hence, a total of 27 conditions will be considered. The three rater scenarios are: (a) all raters have similar variability in their ratings, (b) a large majority of raters exhibits more variability in their ratings, and (c) a small minority of raters exhibits more variability in their ratings. Note that the choices of conditions are intended to reflect operational settings and can be tailored to the assessment context in which LRN examinations are administered. The estimated variance components and score reliability produced by the *rater* and *rating* methods will be evaluated against the true parameters with respect to average bias and root mean square error (RMSE) over 1,000 replications for each condition.

Empirical Study

Operational data from the speaking component of an LRN examination can be used in the empirical study. Based on the results of the simulation study, the method that yields more precise estimates will be applied to estimate variance components of operational data. These estimates will then be used to compute inter-rater reliability and standard error of measurement. Practical recommendations regarding the number of trained raters will be discussed in light of acceptable reliability and reasonable measurement errors.

Although an empirical study by Lee and Kantor (2005) has shown that reliability estimates were similar based on either the *rater* or *rating* method, baseline comparison with true parameters is not possible in empirical research. The proposed study builds on this line of empirical research and expands the scope via Monte Carlo simulations. Specifically, it demonstrates how methodological approaches to be applied to empirical research can be informed by simulation research.

References

Bachman, L. F. & Palmer, A. S. (1996). *Language testing in practice*. Oxford, UK: 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*, 238-257.

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

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

Chapelle, C. A., Enright, M. K., & Jamieson, J. (Eds.) (2008). *Building a validity argument for the Test of English as a Foreign Language.* London : Routledge.

Chiu, C. W. T., & Wolfe, E. W. (2002). A method for analyzing sparse data matrices in the generalizability theory framework. *Applied Psychological Measurement, 26*, 321-338.

Fulcher G. (2003). *Testing second language speaking.* London: Longman/Pearson.

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

Lee, Y.-W., & Kantor, R. (2005). *Dependability of new ESL writing test scores: Evaluating prototype tasks and alternative rating schemes* (TOEFL Report MS-31, RR-05-14). Retrieved from ETS® Monograph Series website: http://www.ets.org/research/policy_research_reports/rr-05-14_toefl-ms-31

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.

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