The Development of an Efficient Teacher-Rater Training Program for Enhancing Rater Reliability

So Young Jang (Honam University)

I. Purpose of This Study

The primary purposes of this study are to develop an efficient training model at the University of Illinois at Urbana-Champaign. This study focused on development of systematic training program for the essay raters. Most of the recent research about rater reliability has focused either on a discrete single aspect by providing a single statistical index, or on one aspect of raters' individual backgrounds, such as raters background information, interaction between rater and task, However, It is not easy to find meaningful implications from these studies for the actual improvement of rater reliability.

It is suggested by many studies that developing a systematic training program through iterative evaluation could be a way to reduce the variability in rating among raters (Choi, 2002; Shin, 2001; Shohamy, Gordon, & Kraemer, 1992; Weigle, 1998). In spite of this suggestion, theoretical frameworks and practical guidance for rater training have had little open discussion, having been handled primarily as an internal practice of individual testing agencies, despite the demand for systematic training programs for enhancing rater reliability.

It seems that standardization would be an effort to see scoring problems as a matter of educational system rather than individual responsibility. This study proposes that a rater training program can be standardized by accomplishing innovative systematic changes that consider the relevant literature, the characteristics of test instrument, the test procedure, and contextual effects such as the concerns of the stake holders (Fulcher & Davidson, 2007).

Three major theories (evaluation theory, training theory, and measurement theory) served for designing the standardized rater training program. First, standardization can be achieved and evaluated by following modified Lynch's program evaluation model (1996; 2003) to formulate a basic framework of standardization, which includes the entire evaluation process from needs analysis to feedback system on the basis of the final product of the evaluation. On the basis of training theory analysis, the most appropriate training materials and methods were created in cooperation with the trainer and staff of a language program. For the development of a systematic training model to fit the rating context, theoretical training models, specific goals, and methods of the training program are defined for a particular training program (Borman, 1977; 1978; 1979 Phillips, 1997; Rothwell & Kazanas, 2004; Waagen, 2006). Finally, measurement theory contributed to evaluating the program's effectiveness and individual raters'

performances. Some issues of rater reliability, measurement, and rating validity were discussed in combination with the effects of the rater training program.

II. Methodology

The data for this study were collected from the ESL Placement Test (EPT) at the University of Illinois at Urbana-Champaign (UIUC) from July, 2009 to January, 2010. This study utilized a modified version of Lynch's program evaluation model (1996; 2003) to collect evidence from different sources, including data drawn from the entire evaluation process ranging from needs analysis to a feedback system based on the final product of the evaluation.

Mixed methods were proposed for the data analysis. Diverse perspectives can provide a better measure of training effectiveness, and were achieved by combining the results of both quantitative and qualitative approaches. Quantitative data analysis was proposed for analyzing the surveys, and the rating corpus. Qualitative and document analysis were also essential for analyzing relevant training materials and workshop observation as well as exploring the degree of change in the perceptions of the raters.

III. Findings and Implications

The EPT training program was innovatively reorganized and upgraded. The trainer integrated and updated the training materials to fit the current rating context. The training program focused on enhancing rater consistency and accuracy, as well as rating validity by reducing systematic errors. The training program was designed so that raters could perform evidence-based judgments and solve their rating problems. The workshop lecture provided substantial information, and raters learned how to accurately observe the essay for the appropriate information to match evidence from the rating scale descriptors, and finally how to make a final decision.

The results of this study provide educational implications for language testing. The salient value of this study is the collaboration with stakeholders in a test administration situation. Raters' concerns and challenges were clearly identified, shared, and resolved with the practitioners (the trainer).

Reference

Borman, W. C. (1977). Consistency of rating accuracy and rating errors in the judgment of human performance. Organizational Behaviour and Human Performance, 20, 238-252.

Borman, W. C. (1978). Exploring upper limits of reliability and validity in job performance ratings. Journal of Applied Psychology, 63(2), 135-144.

Borman, W. C. (1979). Format and training effects on rating accuracy and rater errors. Journal of Applied Psychology, 64(4), 410-421.

Choi, Y. H. (2002). FACETS Analysis of effects of rater training on secondary school English teachers''scoring English writing. Journal of the Applied Linguistics Association of Korea. 18(1), 257-292.

Fulcher, G., & Davidson, F. (2007). Language Testing and Assessment. New York, NY: Routledge

Lynch, B.K. (1996). Language Program Evaluation. Cambridge, UK: Cambridge University Press.

Lynch, B.K. (2003). Language Assessment and program evaluation. Edinburgh, Scotland: Edinburgh University Press.

Phillips, J. (1997). Handbook of Training Evaluation and Measurement Methods. Houston, TX: Gulf Publishing Company.

Rothwell, W.J., & Kazanas, H. C. (2004). Mastering the Instructional Design Process. San Francisco, CA: Pfeiffer.

Shin, D. I. (2001). Exploring rating patterns with Rasch measurement techniques: Iimplications for training. Foreign Languages Education, 8(1), 249-272.

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

Waagen, A.K. (2006). Infoline Guide to Training Evaluation. Alexandria, VA: ASTD Press.

Weigle, S. C. (1998). Using FACETS to model rater training effects. Language Testing, 15(2), 263-287.

Keywords: Teacher-rater Training, Designing rater training program
Applicable Languages: English, Korean
Applicable Level: Tertiary, Teacher Education

Biodata: The author received her Ph.D. in the Educational Psychology at the University of Illinois at Urbana-Champaign. Her research and teaching interests are rater reliability, the development of rater training program, the development of test items for speaking and writing tests, and educational measurement.