Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Assessing Writing 13 (2008) 201–218
ASSESSING WRITING

How accurate are ESL students' holistic writing scores on large-scale assessments?—A generalizability theory approach

Jinyan Huang [*]

College of Education, Academic Complex, 329E, Niagara University, Niagara, NY 14109, United States

Abstract

Using generalizability theory, this study examined both the rating variability and reliability of ESL students' writing in the provincial English examinations in Canada. Three years' data were used in order to complete the analyses and examine the stability of the results. The major research question that guided this study was: Are there any differences between the rating variability and reliability of the writing scores assigned to ESL students and to Native English (NE) students in the writing components of the provincial examinations across three years? A series of generalizability studies and decision studies was conducted. Results showed that differences in score variation did exist between ESL and NE students when adjudicated scores were used. First, there was a large effect for both language group and person within language-by-task interaction. Second, the unwanted residual variance component was significantly larger for ESL students than for NE students in all three years. Finally, the desired variance associated with the object of measurement was significantly smaller for ESL students than for NE students in one year. Consequently, the observed generalizability coefficient for ESL students was significantly lower than that for NE students in that year. These findings raise a potential question about the fairness of the writing scores assigned to ESL students.
© 2008 Published by Elsevier Ltd.

Keywords: Writing assessment; Generalizability theory; Rating variability; Rating reliability; ESL writing

1. Introduction

The assessment of writing has long been considered a problematic area for educational assessment professionals. As stated by Speck and Jones (1998), "there are more problems than

[*] Tel.: +1 716 286 8259.
E-mail address: jhuang@niagara.edu.

1075-2935/$ – see front matter © 2008 Published by Elsevier Ltd.
doi:10.1016/j.asw.2008.10.002

solutions—problems of inter-grader reliability, single-grader consistency, and ultimate accountability for the grades we assign" (p. 17). Variation among and within raters' rating of students' writing contributes to measurement error and thus may threaten the fairness of the assessment of writing (Popham, 1990). Due to the different linguistic and cultural backgrounds of English-as-a-second-language (ESL) students, the assessment of their English writing becomes even more problematic (Connor-Linton, 1995; Hamp-Lyons, 1991; Sakyi, 2000). On the one hand, many factors affect ESL students' writing, including their English proficiency, mother tongue, home culture, and style of written communication (Hinkel, 2003; Yang, 2001). In rating ESL students' writing, raters may differentially consider these factors, and empirical studies have found differences in rater behavior for ESL writing assessments (Bachman, 2000). A number of studies indicate that rater and task as factors affect the assessment of ESL writing. For example, rater background, mother tongue, previous experience, amount of prior training, and types and difficulty of writing tasks have been found to affect the rating of the written responses of ESL students (Santos, 1988; Weigle, 1999). The impact of these factors leads to questions about the accuracy, precision and ultimately, the fairness of the scores obtained from the ratings of written work produced by ESL students.

Increasingly, writing-proficiency standards are being established for both secondary school and university students in North America regardless of students' native languages (Johnson, Penny & Gordon, 2000). Within this context, ESL students have to compete with native English (NE) students in writing. Like NE students, ESL students are expected to successfully demonstrate their ability to write English compositions or complete high-stakes essay examinations (Wiggins, 1993). However, research shows that ESL students face considerable challenges passing these institutional or provincial/state competency examinations of writing (Thompson, 1990). Further these difficulties may be due to more than language deficiencies. As an example, rating inconsistency could be one reason for ESL students' failure and poor performance on these writing examinations. Previous studies have found that raters with different teaching experience assign different scores to the same piece of ESL writing (Hamp-Lyons, 1996; Vaughan, 1991).

It is believed that rating consistency or reliability is essential to sound performance assessment practice, although this presumption has been challenged by some scholars (Moss, 1994). In the context of writing assessment, there may exist unwanted variations in scores due to variations among raters and within raters (Bachman, 1990; Johnson et al., 2000). Both of these variations are problematic as they adversely affect the reliability of the scores assigned to students. Rating reliability indicates the precision of the rating of students' writing, which is related to fairness for test-takers (Johnson et al., 2000). Therefore, rating reliability should be treated as a cornerstone of sound performance assessment.

Classical test theory (*CTT*) is most commonly used as a theoretical framework for the detection of rater variation and estimating reliability in performance assessment situations. However, generalizability (*G*-) theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) is a more powerful approach than *CTT* for the detection of rater variation and estimating reliability (Shavelson, Baxter, & Gao, 1993). *G*-theory extends the framework of *CTT* in order to take into account the multiple sources of variability that can have an effect on test scores. While *CTT* provides a single estimate of error, *G*-theory can be used to identify not only multiple sources of error but also the impact of these sources of error on the overall accuracy (Shavelson & Webb, 1991). Through *generalizability* (*G*-) and *decision* (*D*-) studies, researchers can evaluate the relative importance of various sources of measurement error and interpret score reliability from both norm- and criterion-referenced perspectives. Thus *G*-theory provides a comprehensive conceptual framework and methodology for analyzing more than one measurement facet (factor) simultaneously in investigations of