Assessing Writing 38 (2018) 21–36

Contents lists available at ScienceDirect

Assessing Writing

journal homepage: www.elsevier.com/locate/asw

Integrated writing scores based on holistic and multi-trait scales: A generalizability analysis

Renka Ohta[a,*], Lia M. Plakans[b], Atta Gebril[c]

[a] *University of Iowa, Department of Teaching and Learning, N259 Lindquist Hall, Iowa City, IA, 52242, USA*
[b] *University of Iowa, Department of Teaching and Learning, N274 Lindquist Hall, Iowa City, IA, 52242, USA*
[c] *The American University in Cairo, Department of Applied Linguistics, P.O. Box 71, 11835, New Cairo, Egypt*

ARTICLE INFO

Keywords:
L2 writing assessment
Reading-to-write
Integrated tasks
Rating scales
Score reliability
Generalizability theory

ABSTRACT

Scoring reading-to-write (RTW) tasks is known to be more challenging than independent tasks given that raters should attend to features of source use, in addition to other writing skills. Holistic scales have been traditionally used with this task type; however, analytic scales have recently received increasing attention. While research has looked into score generalizability of RTW tasks, few studies have addressed the impact of rating scales on RTW score reliability. For this purpose, the current study compares score reliability from both holistic and multi-trait rating scales. Following a generalizability theory approach, five raters scored 60 essays written by EFL university students using both holistic and multi-trait scoring methods. The results indicated that scores obtained based on the multi-trait rubric were found to be more reliable than those obtained from a holistic rubric. The results also showed that source use was the most reliable feature among all dimensions of the multi-trait rubric. The study results provide sufficient evidence for the use of multi-trait rating scales in the context of scoring RTW tasks. This outcome could encourage more writing instructors to use these scales in their assessment practices given the detailed information they provide about writers' performance on RTW tasks.

1. Introduction

The notion of an integrated reading-to-write (RTW) task has been well-defined by many scholars in the fields of L1 and L2 acquisition worldwide (Eisterhold, 1990; Grabe, 2001; Hirvela, 2004, 2016). These tasks are seen to elicit writing that provides information on writing skills as well as the students' ability to write from text-based content, a common activity in academic settings. As language and testing programs increasingly incorporate performance assessments that combine writing with other skills, critical aspects of test quality, such as score fidelity, require careful attention, as scores provide valuable feedback to writers and to test users, including teachers. It is important to understand scoring of integrated tasks to assure appropriateness and accuracy of inferences test users make based on writing assessment results (Bachman, Lynch, & Mason, 1995; Weigle, 2002). One line of research in this area seeks to explain the relative contribution of different conditions of measurement (usually called facets) to test scores. Namely, researchers examine potential sources of error contributing to the writing score variance, including rater variability, rating scale, and the writing task (East, 2009; Schoonen, 2005). This investigation is vital in writing assessment as tasks require mediation to assign a score or level to a performance, often through a rating scale.

Rating scales for independent writing tasks have received ample attention in the literature (e.g., Barkaoui, 2007; Schoonen,

[*] Corresponding author.
E-mail addresses: renka-ohta@uiowa.edu (R. Ohta), lia-plakans@uiowa.edu (L.M. Plakans), agebril@aucegypt.edu (A. Gebril).

https://doi.org/10.1016/j.asw.2018.08.001
Received 25 January 2018; Received in revised form 7 May 2018; Accepted 7 August 2018

2005); however, the results cannot be directly applied to RTW score reliability given the unique nature of this task type. In RTW, reading has been found to affect writing: (1) by providing content to enhance background knowledge (Weigle, 2004), (2) by offering linguistic support for writers, such as vocabulary and grammar (Esmaeili, 2002; Plakans & Gebril, 2012), and (3) by modeling writing structure (Leki & Carson, 1997; Plakans & Gebril, 2012). Since reading impacts RTW task performance, rating scales for these tasks should differ from those used with independent writing, and with this difference comes the need to revisit the reliability of RTW scales.

A number of recent studies have investigated issues of integrated assessment rating scale development (Chan, Inoue, & Taylor, 2015; Ewert & Shin, 2015), rater behavior (Gebril & Plakans, 2014), and rater accuracy (Wang, Engelhard, Raczynski, Song, & Wolfe, 2017). Ewert and Shin (2015) traced the development of an empirically-based binary scale for a RTW task involving four teachers. Their interactions uncovered challenges in understanding the nature of a RTW task, fitting the scale with the curriculum, and training raters to use it. In a similar project, Chan et al. (2015) described their process in developing a RTW multi-trait rubric based on expert judgment, a questionnaire, rater feedback, and textual analysis. Both of these studies provide blueprints for developing rating scales for RTW tasks and clarify the difficulties inherent in this process due to the synthesis of reading and writing in these tasks, which sets them apart from independent writing.

Several other studies have explored raters' process of scoring RTW tasks given the additional burden of judging the use of source materials in writing. In a study of scoring integrated tasks, Gebril and Plakans (2014) found that raters attended to seven features in RTW specifically related to source use, including accuracy, relevancy, adequacy, and effectiveness of source use; clarity of source information; appropriateness of textual borrowing strategies; and overuse of source materials. Wang et al. (2017) also considered raters' perceptions and accuracy when assessing integrated writing texts that were identified as "difficult-to-score." Essay focus, textual borrowing, and idea development were found to be more problematic than language use and conventions. These studies of raters and rating scales provide a foundation for understanding aspects of reliability in integrated assessment. More investigation, however, is needed to inform test developers and users on the selection of scoring instruments for RTW assessments.

2. Research review

2.1. Multi-trait and holistic rating scales

Various types of rating scales have been used in second language writing (L2) assessment, but holistic and analytic are the most common (Ebel & Frisbie, 1991). Raters who use holistic scoring focus on communicative aspects of writing or the overall message writers convey (Cooper, 1977; Weigle, 2002) and assign only one score of overall writing quality (Ebel & Frisbie, 1991; Goulden, 1994; Plakans & Gebril, 2015). Multi-trait scales[1] require raters to assign separate scores to particular dimensions of writing, such as organization, language use, and mechanics. When scoring, raters attend to a single writing feature at a time. Multi-trait scales are said to account for "unevenness of quality in the writing" (Hamp-Lyons, 1991, p. 253).

Both scales come with their unique advantages and disadvantages. Holistic scoring is cost- and time-effective; it is suitable for large-scale, high-stakes testing situations in which speedy scoring and overall proficiency of test takers are of primary importance. Multi-trait scoring provides more diagnostic information about the quality of writing, such as strengths and weaknesses (Hamp-Lyons, 1991, 1995; Weigle, 2002) and allows raters to justify final scores (Ebel & Frisbie, 1991; Goulden, 1994). One drawback of holistic scoring is the limited amount of information available for test users, which is problematic in the L2 writing classroom. Multi-trait scales capture the multidimensionality of the writing construct (Hamp-Lyons, 1991, 1995; Weigle, 2002; Wiseman, 2012); however, more time and money are spent on rater training (Lee, Gentile, & Kantor, 2010) and scoring multiple features of a single piece of writing (Weigle, 2002).

Despite the widespread use of two scoring methods, systematic rating scale selection is often neglected in evaluating writing tasks (Bacha, 2001; Ghanbari, Barati, & Moinzadeh, 2012). Becker (2010) investigated types of rating scales used in intensive English programs (IEP) in the U.S. and found that holistic scales were largely preferred over analytic scales. According to the results of Becker's study, teachers found holistic scales less intimidating and more efficient. In reporting the development and refinement of a writing assessment, Bruce and Hamp-Lyons (2015) found that the first iteration of the scale was too complex for raters to score efficiently. Yet, this complex version of the scale was preferred by teachers for giving feedback to their students. It is important to understand the impact of preferring one scale over the other on score reliability so that both scales can be legitimately used depending on the purpose and use of the test.

2.2. Research comparing holistic and multi-trait scores

The empirical research examining the impact of scoring methods on the reliability of writing scores has a long and international history. To our knowledge, the earliest studies include Hartog and Rhodes (1936), Cast (1939, 1940), and Britton, Martin, and Rosen (1966), who investigated the markings of English essays in England using multiple scoring methods, including an impressionistic and a prescribed rating scheme. Researchers in the U.S. context also pursued similar research agendas (e.g., Bauer, 1982; Follman & Anderson, 1967; Veal & Hudson, 1983). Previous research has shown that holistic and multi-trait scores are different in many aspects. More recently and with a focus on L2 writing, Carr (2000) investigated the effect of changing the scales from analytic to holistic on

[1] We use 'multi-trait' and 'analytic' interchangeably in this article.