Contents lists available at ScienceDirect

Assessing Writing

Examining instructors' conceptualizations and challenges in designing a data-driven rating scale for a reading-to-write task

Doreen Ewert[a,*], Sun-Young Shin[b]

[a] University of San Francisco, 2130 Fulton Street, Kalmanovitz Hall 205, San Francisco, CA 94117-1080, USA
[b] Indiana University, 1021 East 3rd Street, Bloomington, IN 47405, USA

ARTICLE INFO

Article history:
Received 5 January 2015
Received in revised form 9 June 2015
Accepted 15 June 2015
Available online 5 August 2015

Keywords:
Second language writing
Rubrics
EBB
Rating
Reading-to-write task
Assessment

ABSTRACT

Integrated reading-to-write (RTW) tasks have increasingly taken the place of independent writing-only tasks in assessing academic literacy; however, previous research has rarely investigated the development and use of rating scales to interpret and score test takers' performance on such tasks. This study investigated how four highly experienced ESL instructors developed an empirically derived, binary choice, boundary definition (EBB) rating scale. EBB scales are known to be reliable and effective for assessing specific writing tasks administered for a single population. Nonetheless, evidence suggests that factors outside the curriculum also influence the criteria which shape an EBB scale and thus final placement scores. Analysis of the recorded deliberations provides evidence of instructors' conceptualizations of reading, writing, and language in the RTW task although each is not equally transparent in the EBB rating scale developed. Understanding the task and the curriculum as well as considering the future training of raters were additional challenges in designing this EBB scale. Despite such challenges, an EBB rating scale has potential to help us better understand the relative contribution of hybrid constructs to the overall quality of RTW task performance and to enhance the linkages among teaching, rating, and future rater-training.

1. Introduction

Given the fact that academic writing is usually done in response to source texts, reading-to-write (RTW) tasks have become a regular feature of assessment practices for determining the proficiency or placement of learners interested in academic pursuits (Carson, 2001; Feak & Dobson, 1996; Horowitz, 1991; Leki & Carson, 1997; Weigle, 2004). This shift to a direct measure of a typical college task can also be seen in the current format of high-stakes tests, such as the Subject A Exam in California, the Canadian English Language Assessment, or the TOEFL iBT, which include a variety of integrated tasks to align with the actual uses of language in the academy. Among these integrated tasks is the reading-to-write (RTW) task.

The content validity of the RTW task is enhanced when it is aligned with a language program curriculum which intentionally combines reading and writing skill development in its courses (Ewert, 2011; Cumming, 2013; Grabe, 2003; Wolfersberger, 2013). Furthermore, a RTW task provides test-takers with source content that reduces the role of students'

* Corresponding author. Tel.: +1 415 422 2607; fax: +1 415 422 2353.
E-mail addresses: dewert@usfca.edu (D. Ewert), shin36@indiana.edu (S.-Y. Shin).

http://dx.doi.org/10.1016/j.asw.2015.06.001
1075-2935/© 2015 Elsevier Inc. All rights reserved.

dissimilar previous knowledge, which can directly affect writing scores (Gebril, 2009; Lewkowicz, 1994; Read, 1990; Weigle, 2004).

However, there are still unanswered questions regarding how to define and operationalize the construct of the RTW task into the rating criteria (Cumming, 2013; Yu, 2013), which ultimately act as the test construct (Knoch, 2011). Although the RTW task may reflect an independent construct that is different from the sum of reading and writing abilities (Asención, 2008), writing has typically been found to be a stronger predictor of RTW task performance than reading (Gebril, 2009; Lewkowicz, 1994; Watanabe, 2001). In a RTW task for second language learners, the raters must focus not only on the test takers' writing ability and reading comprehension as demonstrated through uses of the source text, but also on language ability. Even if these qualities are well-represented in the rating scale, it is still very difficult to achieve rater reliability since it has been found that raters tend to ascribe importance to different qualities of the RTW task across different proficiency levels (Gebril & Plakans, 2014). For these reasons further research into the role of the raters' articulated constructs of RTW task features that differentiate student samples for placement purposes is warranted.

RTW tasks have been rated using both theory-based holistic (Gebril, 2009, 2010; Lee & Kantor, 2005) and performance-based analytic rating scales (Shin & Ewert, 2015). The former type is based a priori on the expected features or components of competent reading and writing from an L1 or idealized L2 perspective. For the latter type, performance samples of test-takers at different levels of proficiency are used to construct the rating scale either by detailed descriptions or in establishing differences between the levels on a scale (Fulcher, Davidson, & Kemp, 2011). In both cases, the rating scales are based on a measurement scale that suggests a linear development of reading, writing, or language use that focuses on the central characteristics of a hypothesized level of performance. Although performance-based rating scales are preferable since they are at least linked to specific tasks for which the rating scales are needed, a great deal of rater variability is still possible when test-takers performances do not represent the central characteristics of the rating scale categories.

Upshur and Turner (1995) have recommended an alternate data-driven approach to rating scale development that addresses the reliability and validity problems of standard rating scales: the empirically derived, binary choice, boundary definition (EBB) scale. The development of an EBB scale begins with the examination of samples of a specific task performance of a single population of test-takers, and not with a theory of how the task can be performed. Through conversations by the rating scale developers during the examination of the actual data, notions emerge regarding how the performances can be clustered to fit the existing levels in a program or letter grades in a course. These notions eventually are formed into yes-no questions that reflect the boundaries of the relevant categories and into which roughly equal numbers of test-takers will fall. There are many factors that affect holistic scoring including the testing environment, the prompts used to generate the language samples, the scoring procedures and rating scales (Hamp-Lyons & Mathias, 1994; Jennings, Fox, Graves, & Shohamy, 1999; Lim, 2010), as well as the development process of the rating scales—an aspect that has not yet received much attention (Turner, 2000; Turner & Upshur, 2002) and not for a RTW task at all. The very nature of the EBB rating scale makes the role of the developers central in the assessment process.

EBB rating scales have been used to assess a variety of specific oral and written language tasks for specific populations (Plakans, 2013; Turner, 2000; Turner & Upshur, 1996, 2002; Upshur & Turner, 1995) with relative success. The EBB rating scales are more closely aligned with curricula, all the categories are used, raters spend less time trying to come to consensus on the minimal number of descriptors the EBB scale development process generates, and inter-rater reliability (IRR) measures have been satisfactory. However, as Turner (2000) points out, the team of scale developers often had differing views that needed to be reconciled, and these conversations definitely affected the content of the scale although Turner and Upshur (2002) found that the scale developers had a minor effect on the ratings.

There is very little research on the content of the deliberations of a scale development team. Turner (2000) provided the first qualitative analysis of the process and discourse stances of the scale developers while developing an EBB rating scale for secondary-level English as a second language (ESL) writing assessment in Quebec. She found that the criteria for the scale were based on the general abilities the development team is focused on measuring at the outset, salient features in the data samples used, and in the lengthy discussions that were necessary to come to consensus on the binary questions the EBB scale requires. Thus, the role of the development team has an impact on the ultimate ratings of the EBB scale. Turner and Upshur (2002) investigated the extent of this impact on the scoring by using three teams of scale developers with two sets of writing samples; Groups A and C used the same samples, and Group B had a different set. They found that each team developed a different rating system in terms of content, but that the IRR within the three scales was high, and between Group A and C the highest. They concluded that while the samples had an effect on the content of the scale, the consistency in the scoring of the raters indicates that there is more than one way to describe text characteristics and still distinguish levels of writing ability. While these findings confirm that scale descriptors are not transparent, the extensive deliberations of a group of scale developers increases the content validity of the rating scale as it is deeply grounded in the shared understandings and experiences in a specific context for a specific task.

More recently, Plakans (2013) reports on the development process of an EBB rating scale to replace an analytic writing assessment tool. Four instructors engaged in the initial rating scale formulation following the guidelines set out in Turner and Upshur (1996). After piloting the scale on 69 writing samples, the raters were only very confident on 22% of the samples, somewhat confident on 75.5%, and not confident on 2.5%. Voiced concerns about some of the wording in the scale content led to revisions. The revised scale continued to undergo review as it was used for placement in subsequent terms. Score distributions, misplacements based on classroom diagnostics, and instructor perceptions of placements and classroom fit were gathered to assess the effectiveness of the scale. Although revisions of the scale have been necessitated by the evolving