Assessing Writing 26 (2015) 5–19

Contents lists available at ScienceDirect

Assessing Writing

Keeping up with the times: Revising and refreshing a rating scale

Jayanti Banerjee[a,*], Xun Yan[b], Mark Chapman[c], Heather Elliott[d]

[a] *Worden Consulting LLC., 115 Worden Avenue, Ann Arbor, MI 48013, USA*
[b] *University of Illinois—Urbana Champaign, 4080 Foreign Languages Building, 707 S. Matthews Avenue, MC-168, Urbana, IL 61801, USA*
[c] *The University of Wisconsin—Madison, WIDA Consortium, 181 Education Building, 1000 Bascom Mall, Madison, WI 53706, USA*
[d] *CaMLA, Argus 1 Building, 535 West William St., Suite 310, Ann Arbor, MI 48103-4978, USA*

ARTICLE INFO

Article history:
Received 2 January 2015
Received in revised form 30 June 2015
Accepted 16 July 2015
Available online 6 August 2015

Keywords:
Rating scale design
Scale validation
Corpora
Discriminant function analysis

ABSTRACT

In performance-based writing assessment, regular monitoring and modification of the rating scale is essential to ensure reliable test scores and valid score inferences. However, the development and modification of rating scales (particularly writing scales) is rarely discussed in language assessment literature. The few studies documenting the scale development process have derived the rating scale from analyzing one or two data sources: expert intuition, rater discussion, and/or real performance.

This study reports on the review and revision of a rating scale for the writing section of a large-scale, advanced-level English language proficiency examination. Specifically, this study first identified from literature, the features of written text that tend to reliably distinguish between essays across levels of proficiency. Next, using corpus-based tools, 796 essays were analyzed for text features that predict writing proficiency levels. Lastly, rater discussions were analyzed to identify components of the existing scale that raters found helpful for assigning scores. Based on these findings, a new rating scale has been prepared. The results of this work demonstrate the benefits of triangulating information from writing research, rater discussions, and real performances in rating scale design.

1. Introduction

In the standardized assessment of writing, rating scale development is a ubiquitous activity. Regular monitoring and modification of the rating scale is also essential to ensure reliable test scores and valid score inferences. However, reports of scale development or revision are rare in language assessment literature. This presents a gap in our discussions of the challenges and opportunities presented during scale development and revision. Of the studies available, most describe the development of speaking assessment scales (e.g. Ducasse, 2009; Upshur & Turner, 1999; Fulcher, Davidson, & Kemp, 2011); there are relatively few studies that address the development of writing scales (Knoch, 2011; Lim, 2012; Sasaki & Hirose, 1999). That said, the scale development process for speaking and writing performances is largely similar and both are therefore relevant for the work presented here.

* Corresponding author. Tel.: +1 7346602767.
E-mail address: j.v.banerjee@gmail.com (J. Banerjee).

http://dx.doi.org/10.1016/j.asw.2015.07.001
1075-2935/© 2015 Elsevier Inc. All rights reserved.

Fulcher et al. (2011) describe the two most common approaches to constructing a rating scale: the measurement-driven, and performance data-driven approaches. The measurement-driven approach starts with the level descriptors. It focuses on the clarity of the descriptors and thus the usability of the rating scale. It also relies on the intuition of experts in language teaching and assessment (e.g., theorists, teachers, or raters) to develop the rating criteria (Hamp-Lyons, 1991). This approach is by far the most commonly used in scale development. However, views on the appropriateness of this approach are mixed. The criticisms of the approach include claims that the resulting scales can lack precision, specificity, and scalability (Fulcher et al., 2011). As the descriptors are often written in impressionistic, abstract, or relativistic language, the distinction between performances across score levels tends to be subjective or less consistent across raters (Knoch, 2009). Concerns have also been raised about the representativeness of the rating scales (Mickan, 2003; Upshur & Turner, 1995). Additionally, intuitively developed scales have been criticized for having descriptors that are inconsistent with theories of L2 development (Turner & Upshur, 2002). The involvement of expert raters in scale development tends to improve the usability of the scale compared with those derived directly from theory in a top-down fashion (Lowe, 1986). However, the intuitive nature of the measurement-driven approach requires no analysis of real performance prior to generating descriptors. This makes the resultant rating scales dependent upon post-hoc quantitative or qualitative analysis to ensure reliability of the descriptors and validity of the score inferences.

The performance data-driven approach, on the other hand, derives rating scales through analyzing real language performances. This approach starts with performances, and identifies traits or features that characterize and discriminate written texts or writers across proficiency levels. There are two sub-approaches within the performance data-driven approach (Council of Europe, 2001, p. 207): qualitative and quantitative methods. The qualitative method pre-tests the effectiveness of descriptors derived from the measurement-driven approach through detailed analysis of a small number of test performances. The quantitative method quantifies and cross-validates the qualitative evidence on a larger scale. The two methods are clearly complementary (Lim, 2012), and are thus recommended to be used in combination. Unlike the post-hoc reliability or validity analysis in the measurement-driven approach, the analyses in the performance data-driven approach are primarily exploratory in nature. That is, the analyses of performance data precede the development of the scale and are not aimed at confirming a pre-determined set of features. The advantage of this approach lies in the resulting scale's reflection of real performances. However, data-based analysis tends to be time consuming. Additionally, in a completely data-driven approach, especially when using corpus-based tools, the data tend to generate linguistic constructs that either bear complex mathematical formulae or become extremely difficult to operationalize by human raters (Fulcher, 2003). The level descriptors would need to be carefully written in order to ensure that the linguistic features are accessible to examiners. Additionally, rater training would need to be carefully structured so that divided and yet simultaneous attention to individual criteria is possible but not over-taxing for raters in real-time rating.

In addition to the aforementioned approaches, the literature on scale development has called for more theory-based practices in scale development (e.g., Fulcher, 1987; Knoch, 2011; McNamara, 2002). Lantolf and Frawley (1985) have argued that a lack of linkage between theories of L2 development and construct representation raises questions about the validity of the rating scale. Despite this there are no records of a scale development process using theory to inform its construction. This is perhaps, as argued by Knoch (2011) and Lantolf and Frawley themselves, due to the lack of a unified theory of L2 development or language proficiency. This makes it difficult to develop rating scales using a theory-based approach.

It appears, therefore, that the most defensible approach to rating scale development and revision would be to adopt an approach that combines our current understanding of the indicators of second language writing development (cf. Wolfe-Quintero, Inagaki, & Kim, 1998), expert intuition, and the empirical analysis of performance data. This is the approach that we have taken in the review and revision of the rating scale for the writing section of a large-scale advanced level English language proficiency examination. We have triangulated three data sources by: reviewing expert intuition and analysis to build a framework of the text features that are expected to predict writing proficiency; using corpus tools to analyze 796 real performances; and analyzing rater discussions during the scoring process.

2. Background to the study

The rating scale under review here is the assessment tool for the writing section of a large-scale English language proficiency examination designed for advanced-level learners, the Examination for the Certificate of Proficiency in English (ECPE). Developed by CaMLA (http://www.cambridgemichigan.org/), the exam comprises four sections, writing, listening, reading, and speaking. The results for each section are reported separately. The writing section is 30 min long and offers test takers a choice of two essay prompts. They choose one and are expected to write at least 300 words. Both prompts require test takers to give their opinion on a statement and to justify that opinion using supporting details or points. Test takers who pass the writing section are considered to be at C2 on the Common European Framework of Reference (CEFR, Council of Europe, 2001). They are able to communicate their ideas fully in clear, smoothly flowing language. They can structure their text logically to present an effective argument and can use grammatical structures and vocabulary flexibly in order to convey precise meaning. As such, the intended construct of the writing section includes breadth and depth of vocabulary knowledge, variety and accuracy of grammatical structures, ability to state and develop an argument, audience awareness, and text organization skills.