Assessing Writing 26 (2015) 51–66

Contents lists available at ScienceDirect

Assessing Writing

Building a better rubric: Mixed methods rubric revision

Gerriet Janssen[a,b,*], Valerie Meier[c], Jonathan Trace[b]

[a] Universidad de los Andes, Colombia
[b] University of Hawai'i, Mānoa, USA
[c] University of California, Santa Barbara, USA

ARTICLE INFO

Article history:
Received 8 January 2015
Received in revised form 6 July 2015
Accepted 16 July 2015
Available online 6 August 2015

Keywords:
Academic writing
Mixed-methods
Profile analysis
Rasch measurement
Rubrics

ABSTRACT

Because rubrics are the foundation of a rater's scoring process, principled rubric use requires systematic review as rubrics are adopted and adapted (Crusan, 2010, p. 72) into different local contexts. However, detailed accounts of rubric adaptations are somewhat rare. This article presents a mixed-methods (Brown, 2015) study assessing the functioning of a well-known rubric (Jacobs, Zinkgraf, Wormuth, Hartfiel, & Hugley 1981, p. 30) according to both Rasch measurement and profile analysis ($n = 524$), which were respectively used to analyze the scale structure and then to describe how well the rubric was classifying examinees. Upon finding that there were concerns about a lack of distinction within the rubric's scale structure, the authors decided to adapt this rubric according to theoretical and empirical criteria. The resulting scale structure was then piloted by two program outsiders and analyzed again according to Rasch measurement, placement being measured by profile analysis ($n = 80$). While the revised rubric can continue to be fine-tuned, this study describes how one research team developed an ongoing rubric analysis, something that these authors recommend be developed more regularly in other contexts that use high-stakes performance assessment.

1. Introduction

Scoring rubrics are important as they articulate the construct to be performed and measured. Rubrics "help explain terms and clarify expectations" (Crusan, 2010, p. 43). This is to say, the principled choice and use of rubrics is vital, as rubrics optimally link the task, the constructs developed by the task, and the assessment of these constructs. Weigle (2002) describes how the scoring process using rubrics can be particularly "critical because the score is ultimately what will be used in making decisions and inferences about writers" (p. 108). Rubrics can also help mitigate the long-recognized problem of rater variability (cf. Bachman et al., 1995; McNamara, 1996).

Recognizing the importance of rubrics, local program developers – when developing the *Inglés para Doctorados* (IPD; English for Ph.D. students) program and the corresponding IPD Placement Exam used to classify students into the program's courses – decided to use the analytic rubric developed by Jacobs, Zinkgraf, Wormuth, Hartfiel, and Hugley (1981, p. 30) for use with the performance writing component of the placement exam (Janssen et al., 2011). This rubric was adopted because of the strong construct validity it had in terms of proposed course goals and because Weigle (2002, p. 115), had characterized this rubric as being "one of the best known and most widely used analytic scales in ESL". Indeed, the Jacobs et al. (1981)

* Corresponding author at: Universidad de los Andes Depto. de Lenguajes y Estudios Socioculturales Cra 1 No. 18A-12 Bogotá, Colombia. Tel.: +57 1 339 4949x3248.
E-mail address: gjanssen@uniandes.edu.co (G. Janssen).

http://dx.doi.org/10.1016/j.asw.2015.07.002
1075-2935/

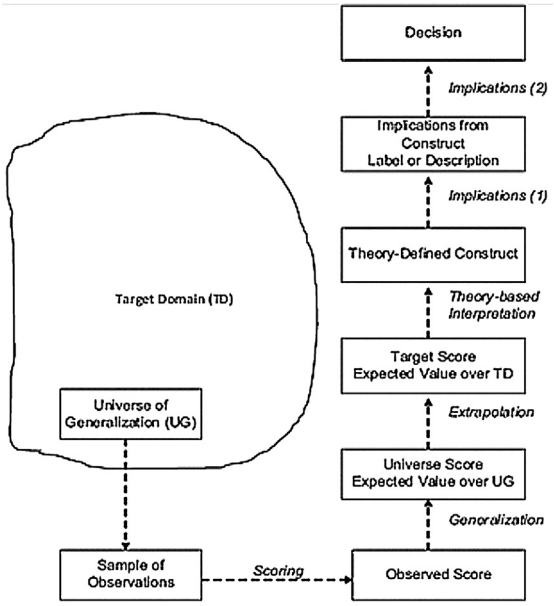

Fig. 1. Chapelle's (2012, p. 23) conceptualization of Kane's IUA.

rubric, in its original or modified form, has been used with some frequency (cf., Brown & Kondo-Brown, 2012; Delaney, 2009; East, 2009; Polio, 2001; Winke & Lim, 2015).

With time, several concerns arose concerning the IPD Placement Exam and its different uses. Though originally designed for use as a placement instrument, Ph.D. program directors began using different course level placements as one admission criterion for entrance into the university's Ph.D. programs. With this unforeseen high-stakes use, program developers began to intensively study different aspects of the exam's validation argument. Of relevance to this current study, in Janssen and Meier (2012) we first realized that the rubric chosen to score the performance writing section, while reliable, was not performing as expected. Indeed, the reliability reported for the Rasch model was 0.99, yet other indicators such as threshold distances (i.e., increments in difficulty) pointed to problems within the scoring bands of the rubric itself: increasing scores were not consistently representative of increased examinee ability (Janssen & Meier, 2012; Meier, 2013). Furthermore, interviews with exam raters, though generally positive in terms of the constructs the rubric represents, revealed other concerns with the rubric, specifically in relation to its ease of use when scoring. Thus, the current study: (a) considers the function of the original Jacobs et al. (1981) rubric; (b) proposes a reformulated rubric that addresses the scoring band problems and answers the raters' call for added simplicity; and (c) analyzes the functioning of this revised rubric.

2. Literature review and research questions

2.1. Validation

Following the work of Bachman and Palmer (2010), Chapelle (2008, 2012), and Kane (2006, 2013), IPD program developers have been building a validation argument for the uses of the IPD Placement Exam using an argument-based validity framework. Kane (2013) presents six sequential inferences that are typically addressed in the interpretation-use arguments (IUAs) for placement exams: scoring, generalization, extrapolation, theoretical, and two levels of implications. These inferences should be defended in the IUAs of most tests, though part of what makes the argument-based approach to validation so powerful is that the arguments claimed within each IUA should adjust themselves to the interpretations and uses found within the specific assessment context. Chapelle's (2012) helpful depiction of Kane's IUA has been included as Fig. 1.

In this paper, we focus on elements of the scoring inferences within the IUA. The scoring inference focuses on the scoring procedures and the application of these procedures to ensure that they are appropriate, accurate, and consistent (Kane, 2006, pp. 24, 34; Kane, 2013, p. 25). Clauser (2000) provides an in-depth description of several important components of the scoring inference of appropriacy that should be evidenced. Three key components to demonstrate include determining: (a) if the constructs developed within the rubric are appropriate to the larger construct being evaluated in this exam section; (b) whether the criteria used for evaluation are appropriate; and (c) if these are being applied in an appropriate fashion. The appropriateness of the rubric's constructs and criteria of evaluation can be evaluated by field experts; the appropriateness of application can be judged using Rasch measurement, which provides test developers with a variety of analyses (e.g., bias, fit/misfit, reliability, scale analysis) that can be done to help demonstrate how the test is functioning, and to what