Assessing Writing 27 (2016) 1–10

Contents lists available at ScienceDirect

Assessing Writing

Features of difficult-to-score essays

Edward W. Wolfe [a,*], Tian Song [b,1], Hong Jiao [c,2]

[a] Research & Innovations Network, Pearson, 3974 Roberts Ridge NE, Iowa City, IA 52240, USA
[b] Pearson, 5604 E Galbraith Rd., Cincinnati, OH 45236, USA
[c] University of Maryland, 1230C Benjamin Building, College Park, MD 20742, USA

ARTICLE INFO

Article history:
Received 31 October 2014
Received in revised form 11 June 2015
Accepted 18 June 2015
Available online 7 August 2015

Keywords:
Rater
Scoring
Writing assessment

ABSTRACT

Previous research that has explored potential antecedents of rater effects in essay scoring has focused on a range of contextual variables, such as rater background, rating context, and prompt demand. This study predicts the difficulty of accurately scoring an essay based on that essay's content by utilizing linear regression modeling to measure the association between essay features (e.g., length, lexical diversity, sentence complexity) and raters' ability to assign scores to essays that match those assigned by expert raters. We found that two essay features – essay length and lexical diversity – account for 25% of the variance in ease of scoring measures, and these variables are selected in the predictive modeling whether the essay's true score is included in the equation or not. We suggest potential applications for these results to rater training and monitoring in direct writing assessment scoring projects.

1. Introduction

Direct writing assessments, as they are employed in many educational contexts, require examinees to compose an essay in response to a prompt or stimulus material, sometimes following a sequence of guided prewriting activities. These assessments may be informal, taking place within the classroom and providing formative information regarding instruction, or they may be formal, taking place within a standardized assessment setting and providing information upon which summative and/or accountability decisions are based. Whether the purpose of the assessment is formal or informal, those essays are typically scored by human raters who employ a scoring rubric as a guide to classifying essays into ordered categories intended to indicate increasing levels of writing quality. Those scoring rubrics may focus on and require assignment of several "trait" scores, each of which depicts the quality of a particular aspect of the writing (e.g., mechanics, organization, development, voice), or the rubrics may require the rater to make a holistic judgment by jointly considering all relevant aspects of the essay in arriving at a single score of the overall quality of the writing.

That rating process, because it involves subjective judgments, may, potentially, result in different scores being assigned to the same essay by different raters. In fact, that decision making process is sufficiently subjective to allow fluctuations in a rater's attention and mood, changes in the scoring context, and even variations in the presentation of the essays being scored to cause a single rater to assign different scores to the same essay on different occasions (Shohamy, Gordon, & Kraemer,

* Corresponding author. Tel.: +1 319 321 4633.
E-mail addresses: ed.wolfe@pearson.com (E.W. Wolfe), tian.song@pearson.com (T. Song), hjiao@umd.edu (H. Jiao).
[1] Tel.: +1 734 546 4239.
[2] Tel.: +1 301 405 3627.

http://dx.doi.org/10.1016/j.asw.2015.06.002
1075-2935/

1992). Prior research concerning this decision making process identifies several components of the assessment context may influence the accuracy of rater decision-making. This manuscript reports the results of a study of one of those components, the features of essays, which is associated with disagreements between raters with the intent of providing insights that may inform rater training and monitoring practices in direct writing assessments.

1.1. Terminology

Before discussing the various potential sources of rater disagreements, we define important terms that are relevant to our focus. The term *interrater reliability* is often used loosely to refer to the degree to which several potential types of measurement error may influence scores assigned by human raters, and we avoid that term due to the potential confusion that it may introduce. We prefer the specificity of the terms *interrater agreement* and *rater accuracy* (sometimes referred to as *rater validity*). Interrater agreement typically refers to the degree to which a rater assigns scores to a particular set of examinee responses that are consistent with scores assigned to those responses by other raters. Rater accuracy, on the other hand, refers to the degree to which a rater assigns scores to a set of responses that match *validity scores* (i.e., scores that are assumed to be accurate scores, typically consensus scores assigned by expert raters). The only difference between rater agreement and rater accuracy is the frame of reference against which we evaluate the rater's performance.

We attribute rater disagreement or rater inaccuracy to patterns within the scores assigned by a particular rater which may have diagnostic utility, which we call *rater effects*. Those patterns exhibit some degree of predictability and may cause the assigned scores for the rater to be consistently high or low (severity or leniency), to be tightly or widely spread (centrality or extremity), or highly consistent or inconsistent (accuracy or inaccuracy) when compared to the target scores. Hence, when rater effects are identified, raters can be provided with information concerning why low interrater agreement or low levels of rater accuracy have occurred and what can be done to correct those errors. Our point is that rater errors can be depicted at either the more general rater agreement/accuracy level or the more diagnostically specific rater effect level, with rater agreement/accuracy being more comprehensive and rater effects being more diagnostically specific. In this manuscript, we focus on the level of rater agreement/accuracy because our purpose is to discover sources of rater errors that may arise due to how raters respond to the content and quality of a particular essay rather than to diagnose the types of errors that raters commit.

1.2. Influences on rater agreement and accuracy

Several conceptual models have been proposed to explain how various components of the rating process and context may contribute to rater disagreement and inaccuracy, and we synthesize some of those models in this section in order to identify the context within which our research is situated. Generally, four broad aspects of the rating process have been proposed as potential antecedents to the emergence of rater effects. First, although it has not been subjected to much formal research, the design of the assessment clearly has the potential to impact the quality of assigned scores. Here, we are referring to decisions that have been made on the part of assessment designers, such as the purpose of the assessment, the administration medium, and the focus of the scoring criteria. One example of research that has focused on the impact of direct writing assessment design on rating quality concerns the comparability and raters' perceptions of essays that are composed in handwriting versus word processor. For example, Powers, Fowles, Farnum, and Ramsey (1994) performed an early study of this topic by asking students to compose essays in both of these composition media and then transcribing each essay to the other medium and finally having raters assign scores to both versions of each essay. Their results indicated that handwritten essays received higher scores, regardless of composition medium. That is, raters, on average, exhibited a leniency effect for handwritten essays. It is worth acknowledging what may seem a potential overlap between assessment design decisions and the next feature that we discuss, response content. Our decision to discuss essay composition medium as an example of assessment design is based on the fact that composition medium in direct writing assessments is typically a decision that is imposed on the examinee by assessment designers, while other characteristics of the response, such as handwriting quality, are more directly under the control of the examinee. Classifying other examples of research relating to assessment design, such as the assessment purpose and the nature of the scoring criteria, would be more straightforward.

Second, the content of the response that raters review, which is the primary focus of the research that we conducted, may also influence the quality of assigned scores. In the context of direct writing assessment, when we refer to response content, we mean the visual appearance of the response (e.g., handwriting quality, font choices, and page layout), textual features (e.g., length, word choice, mechanics), and content included in the response (e.g., author clues, ideas). A great deal of research has been conducted regarding the impact of visual appearance on rating quality, but relatively less research exists concerning the impact of textual features and informational content. We briefly summarize this research in the following section.

Third, rater characteristics have long been posited as potential influences on rating quality (Pula & Huot, 1993). Rater characteristics include rater experiences (e.g., educational, demographic, and professional), stable rater cognitive and affective traits (e.g., temperament, cognitive style), and temporary rater states (e.g., mood). A good bit of research has been conducted regarding the impact of rater experiences (Wolfe, 1997; Meadows & Billington, 2010; Pula & Huot, 1993; Shohamy et al., 1992; Sweedler-Brown, 1985) and stable rater traits (Crisp, 2012; Huot, 1993; Vaughan, 1991) on rating quality. Generally, these studies suggest that the more general the rater background characteristic (e.g., demographics), the less likely it is to