Assessing Writing 33 (2017) 36–47

Contents lists available at ScienceDirect

Assessing Writing

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

# Evaluating rater accuracy and perception for integrated writing assessments using a mixed-methods approach

Jue Wang[a,*], George Engelhard Jr.[a], Kevin Raczynski[a], Tian Song[b], Edward W. Wolfe[c]

[a] The University of Georgia, United States
[b] Pearson, United States
[c] Educational Testing Service, United States

ARTICLE INFO

*Keywords:*
Rater accuracy
Rater perception
Integrated writing assessment
Mixed-methods approach
Rasch measurement theory

ABSTRACT

Integrated writing (IW) assessments underscore the connections between reading comprehension and writing skills. These assessments typically include rater-mediated components. Our study identified IW type essays that are difficult-to-score accurately, and then investigated reasons based on rater perceptions and judgments. Our data based on IW assessments are used as formative assessments designed to provide information on the developing literacy of students. We used a mixed- methods approach with rater accuracy defined quantitatively based on Rasch measurement theory, and a survey-based qualitative method designed to investigate rater perceptions and judgments toward student essays within the context of IW assessments. The quantitative analyses suggest that the essays and raters vary along a continuum designed to represent rating accuracy. The qualitative analyses suggest that raters had inconsistent perceptions toward certain features of essays compared to the experts, such as the amount of textual borrowing, the development of ideas, and the consistency of the focus. The implications of this study for research and practice of IW assessments are discussed.

## 1. Introduction

The use of integrated writing (IW) assessments has been increasing in the United States (Plakans, 2015; Weigle & Montee, 2012). Some prominent examples include assessments developed by PARCC (parcc.pearson.com), Smarter Balanced (www.smarterbalanced.org), the College Board (www.collegeboard.org), and state departments of education. IW assessments usually involve asking students to write an essay based on a given set of reading materials. IW assessments are different from typical writing tasks because IW assessments stress connections between the reading comprehension and writing skills of students that more closely reflect language use in actual academic settings (Weigle & Montee, 2012). Research studies related to IW assessments are emerging and in the spotlight (*e.g.*, Chan, Inoue, & Taylor, 2015; Cumming, 2013; Plakans, Gebril, & Bilki, 2016; Weigle & Montee, 2012).

The validity, reliability, and fairness of the scoring decisions based on the performance ratings remain important with the advent of IW assessments within the context of literacy testing. In a qualitative study on rater perceptions of textual borrowing in IW type assessments, Weigle and Montee (2012) found that raters (a) place different importance on formal conventions and (b) had different attitudes toward paraphrasing text. New challenges are introduced for raters with the IW assessments because raters must evaluate both the written expression of students, as well as their skills in reading in order to incorporate relevant information from provided documents into their essays.

* Corresponding author at: 126C Aderhold Hall, 110 Carlton St Athens, GA 30602, United States.
*E-mail address:* cherish@uga.edu (J. Wang).

http://dx.doi.org/10.1016/j.asw.2017.03.003
Received 23 August 2016; Received in revised form 5 March 2017; Accepted 21 March 2017

Rating accuracy is defined as a comparison between ratings and a corresponding set of scores accepted as the standard for the performance (Sulsky & Balzer, 1988). Sulsky and Balzer (1988) suggested three ways for developing "true scores" (referred to as criterion ratings in this study): (a) using average scores across all raters, (b) using average ratings of previous scaled performance tasks as an anchor file, and (c) using the ratings from a group of expert raters. In this study, we defined the criterion scores as the expert *consensus scores*. Engelhard (1996, 2013) defined rater accuracy as a latent variable that can be measured by the accuracy ratings. In this study, we also view rater accuracy as a latent continuum that can be examined using accuracy data. In other words, scoring tasks are regarded as a test-like activity for raters, and accuracy ratings are obtained to evaluate their scoring proficiency.

It is essential in large-scale writing assessments to train and monitor raters in order to create accurate, consistent, and fair ratings. For direct writing assessments, Raczynski, Cohen, Engelhard, and Lu (2015) demonstrated that some essays are significantly more difficult to be scored accurately than other essays for the professional raters. Wolfe, Song, and Jiao (2016) investigated features of difficult-to-score essays for professional raters, and found that essay length and lexical diversity accounted for 25% of the variances in the difficulty of accurately scoring an essay. Rater training and monitoring related to IW assessments must address the unique scoring challenges associated with these assessments. It is essential that training methods build a consistent frame that presents a cognitive model of accurate scoring for raters. There are several methods used for rater training, and one of the widely used methods is the frame-of-reference training (Bernardin & Buckley, 1981). We used this training method in this study because it is commonly used in conjunction with monitoring rater errors and accuracy (Johnson, Penny, & Gordon, 2009). The detailed training procedures are discussed later.

A mixed-methods approach (Fraenkel, Wallen, & Hyun, 1993) provides a way to examine the relationship between the quantitative ratings by human raters and their perceptions and judgmental processes in depth. In particular, the qualitative component of this study offers the opportunity to learn more about the reasoning used by raters when they provide ratings that are not congruent with the ratings provided by an expert panel. This information holds the potential to improve training for raters who are learning to score accurately, reliably and fairly on IW assessments.

## 2. Purposes

The purposes of this study are to evaluate rater accuracy in scoring essays within the context of IW assessments, and to examine rater perceptions toward essays with various challenges to score accurately. The quantitative analyses target the measurement of difficulty in scoring the essays accurately, and the qualitative data analyses explore the reasons that some essays were difficult for raters to score accurately based on rater justifications of scoring decisions. The following specific research questions were addressed:

1. Is there evidence indicating that some essays are more difficult to score accurately?
2. Based on the justifications that raters provide for their scoring decisions, are there essay features associated with rater perceptions that suggest why some essays are difficult to score accurately?

## 3. Methodology

### 3.1. Participants

Twenty professional raters were randomly selected from the operational rater pool at the Georgia Center for Assessment. There were four male raters (20%) and 16 female raters (80%). The years of experience for the sampled raters had a median of 9 years with a range from 2 to 23, and had a mean of 9.6 with a standard deviation of 5.4.

We recruited a panel of three experts at the Georgia Center for Assessment (two males and one female). These three experts have 6, 7, and 11 years of scoring experience respectively. The defining features of expert raters were that they have a history of accurate scoring of IW assessments, and they conducted the initial range-finding in August 2014 for the specific IW assessment associated with our study. During range-finding, these three experts were involved in creating and selecting the training materials used in this project. This included the scoring rubric and the essays used in training (*i.e.*, training essays).

Grade 7 students in classrooms within two local districts in the Southeastern U.S. were invited to participate in our study. One district was a large urban district and the other was a rural district. These districts were selected because the total number of students from each district was large resulting in a large pool of essays, from which we randomly selected 100 student essays as validity essays to be used in our study.

### 3.2. Instrument

The IW assessment in use was for Grade 7 and developed by the Georgia Center for Assessment. The purpose of this formative assessment was to provide diagnostic feedback on students' developing literacy skills, specifically their skills in reading and writing informational texts. In this assessment, students first read a pair of passages with approximately 500 words per passage, and then responded to three selected-response items, one constructed-response item, and one extended-response item. Of particular interest for this study were the student responses to the extended-response item (*i.e.*, an essay item) which included two reading passages and a prompt. Passage 1 described the causes and effects of water crises in parts of Africa. Passage 2 told the story of a girl who lives in an area of Africa affected by a water crisis. The instructions for the extended-response item were:

*Write an informational essay describing the consequences of the water crisis that some African countries face. In your essay, be sure to:*