Exploring the relationship between textual characteristics and rating quality in rater-mediated writing assessments: An illustration with L1 and L2 writing assessments
Stefanie A. Wind⁎, Catanya Stager, Yogendra J. Patil
The University of Alabama, United States

A R T I C L E   I N F O

Keywords:
Writing assessment
Rating quality
Rasch measurement theory
English language learners

A B S T R A C T

Numerous researchers have explored the degree to which specific textual characteristics of student compositions are associated with high and low ratings, as well as differences in these relationships across subgroups of students (e.g., English language learners). These studies provide insight into rater judgments and the development of writing proficiency. However, the degree to which textual characteristics are associated with the psychometric quality of ratings is relatively unexplored. This study illustrates a procedure for exploring the influence of textual characteristics of essays on rating quality in the context of rater-mediated writing performance assessments in order to gain a more-complete understanding of rating quality. Two illustrative datasets are used that reflect writing assessments for native English speakers and English language learners. The CohMetrix software program was used to obtain measures of textual characteristics, and the Partial Credit model was used to obtain indicators of rating quality. The relationship between essay features and rating quality was explored using correlation and profile analyses. Results suggested that rating quality varies across essays with different features, and the relationship between rating quality and essay features is unique to individual writing assessments. Implications are discussed as they relate to research and practice for rater-mediated writing assessments.

Cooksey et al., 2007
Cooksey, Freebody, & Wyatt-Smith, 2007
Standards for Educational and Psychological Testing

A rater might reserve the highest score level for test takers who provide more information or elaboration than was actually requested. In this situation, test takers who simply follow instructions, or test takers who value succinctness in responses will earn lower scores; thus, characteristics of the individuals have become construct-irrelevant components of the test scores. (p. 56)

In order to ensure fair assessment procedures for all students, it is essential to examine the degree to which rating quality is invariant across essays with different textual characteristics.

In the context of rater-mediated writing assessments, several researchers have explored the relationship between textual characteristics of student essays and the scores assigned to these essays (i.e., judged writing achievement; Breland, Bonner, & Kubota, 1995; Kobrin, Deng, & Shaw, 2007; Kobrin, Deng, & Shaw, 2011). In particular, several recent studies provide evidence of systematic
relationships between text complexity and rater judgment of writing achievement, with higher levels of linguistic complexity and higher word counts positively associated with essay scores (e.g., Yang, Lu, & Weigle, 2015). Similarly, several researchers have explored the relationship between essay length and judged writing achievement, where longer essays are generally associated with higher ratings (e.g., Chodorow & Burstein, 2004; Kobrin et al., 2007).

In research on second-language writing assessment, several researchers have examined the textual characteristics of student compositions across different types of writing tasks. For example, Cummings et al. (2006) examined the textual characteristics of English as a Second Language (ESL) students' compositions on the Test of English as a Foreign Language (TOEFL; "TOEFL iBT: About the Test," n.d.) and observed differences in the textual characteristics of student compositions across different types of writing tasks. Furthermore, these authors observed differences in the textual characteristics of student compositions across judged achievement levels, where students with higher proficiency composed essays that were generally longer, and included more clauses and summaries. Knoch, Macqueen, and O'Hagan (2014) also examined the textual characteristics within TOEFL student compositions related to types of writing tasks and judged proficiency levels. Similar to Cummings et al., Knoch et al. found differences in the textual characteristics of ESL students' compositions across writing tasks, and also observed different patterns of textual characteristics across different types of tasks, as well as across different achievement levels. Similar results were also reported by Banerjee, Franceschina, and Smith (2007), who explored textual characteristics of ESL student compositions across different writing tasks and different levels of judged achievement on the International English Language Testing System (IELTS; "IELTS Exam Format," n.d.). These authors observed differences in students' use of vocabulary across different writing tasks, along with differences in students' use of grammatical markers (e.g. demonstratives such as “this” or “these”) across levels of judged achievement, where students' use of grammatical markers tended to taper off as their judged proficiency level increased.

Despite the large body of research related to relationships between textual characteristics of student compositions and levels of judged writing proficiency, research that has explored relationships between essay features and indicators of the quality of ratings (e.g., rater agreement, rater errors and systematic biases, and rater accuracy) is limited. Information about the degree to which various essay features are associated with rating quality is needed in order to more fully understand the influence of cues in essays that may influence the rating process. Recently, Wolfe, Song, and Jiao (2016) explored correlations between textual characteristics and indicators of rater accuracy in the context of a writing assessment, where rater accuracy was defined using the match between observed scores and scores assigned by an “expert rater”. In this context, the term “expert rater” refers to a rater who is viewed as having more expertise than the operational raters and thus serves as a criterion against which to evaluate operational raters. Essay features were detected using a combination of automated text analysis based on exclusive software, and qualitative coding by the researchers. Using correlation and regression analyses, these authors observed significant positive relationship between lexical diversity and rater accuracy, such that essays with more-diverse language were associated with higher levels of rater accuracy. Building upon Wolfe et al., we examine relationships among textual characteristics that can be obtained from a freely available text analysis program and an indicator of rating quality that does not require “expert raters” (additional raters besides those who participate in operational scoring, who are viewed as experts or criteria against which to evaluate operational raters) to score each essay. Specifically, we use model-data fit statistics for essays to examine rating quality at the essay level (described further below). As a result, the procedures illustrated in this study can be more readily adopted in rater training and monitoring procedures for operational writing assessment programs.

1. Purpose

The primary purpose of this study is to explore the relationship between textual characteristics of essays (i.e., essay features) and the quality of rater judgments (i.e., rating quality) related to individual essays. As a secondary purpose, this study illustrates a method exploring the influence of textual characteristics of essays on rating quality in the context of rater-mediated writing performance assessments. Accordingly, the following research questions guide this study:

1. How do measures of textual characteristics correspond to rating quality indices in rater-mediated writing performance assessments?
2. To what extent are profiles of textual characteristics similar across essays with different levels of judged achievement and different values of fit statistics?
3. To what extent does information about relationships between textual characteristics and rating quality indices provide insight into the psychometric quality of rater-mediated writing assessments beyond traditional (i.e., standalone) rating quality analyses?

This study contributes to research and practice for rater-mediated assessments in several ways. First, whereas most previous research on textual characteristics has examined the relationship between textual characteristics and writing proficiency, we extend research on textual characteristics to include an investigation of their relationship to the quality of ratings that are assigned in rater-mediated writing assessments. In so doing, we also demonstrate this procedure using freely available text analysis tools and rating quality indices that can be calculated without the need for separate criterion or “expert” ratings on each essay besides those that are assigned during operational scoring. Second, we extend previous research on quantitative rating quality indicators to incorporate characteristics of student performances. Examining the influence of these characteristics on rating quality leads to a more complete understanding of the types of student compositions that may contribute to rater errors and other idiosyncrasies in human scoring. Relatedly, researchers and practitioners can implement the procedures illustrated in this study to identify textual characteristics that are difficult for raters to score within writing assessment programs beyond those described in this study. This information can be used...
دریافت فوری متن کامل مقاله

امکان دانلود نسخه تمام متن مقالات انگلیسی
امکان دانلود نسخه ترجمه شده مقالات
پذیرش سفارش ترجمه تخصصی
امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
امکان دانلود رایگان ۲ صفحه اول هر مقاله
امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
دانلود فوری مقاله پس از پرداخت آنلاین
پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات