4-2019

Developing and Examining Validity Evidence for the Writing Rubric to Inform Teacher Educators (WRITE)

Tracey S. Hodges
University of Alabama

Katherine Landau Wright
Boise State University

Stefanie A. Wind
University of Alabama

Sharon D. Matthews
Texas A&M University

Wendi K. Zimmer
Texas A&M University

See next page for additional authors

Publication Information

Hodges, Tracey S.; Wright, Katherine Landau; Wind, Stefanie A.; Matthews, Sharon D.; Zimmer, Wendi K.; and McTigue, Erin. (2019). "Developing and Examining Validity Evidence for the Writing Rubric to Inform Teacher Educators (WRITE)". *Assessing Writing, 40*, 1-13. https://doi.org/10.1016/j.asw.2019.03.001

Authors

Tracey S. Hodges, Katherine Landau Wright, Stefanie A. Wind, Sharon D. Matthews, Wendi K. Zimmer, and Erin McTigue

Developing and Examining Validity Evidence for the Writing Rubric to Inform Teacher Educators (WRITE)

Resubmitted to *Assessing Writing* February 2019

Katherine Landau Wright*
Boise State University
katherinewright@boisestate.edu

Tracey S. Hodges
University of Alabama

Wendi K. Zimmer
Texas A&M University

Erin M. McTigue
University of Stavanger

Abstract

Assessment is an under-researched challenge of writing development, instruction, and teacher preparation. One reason for the lack of research on writing assessment in teacher preparation is that writing achievement is multi-faceted and difficult to measure consistently. Additionally, research has reported that teacher educators and preservice teaches may have limited assessment literacy knowledge. In previous studies, researchers have struggled to provide strong evidence of validity, reliability, and fairness across raters, writing samples, and rubric items. In the present study, we fill several gaps in the research literature by developing a rubric, the Writing Rubric to Inform Teacher Educators (WRITE), which utilizes a structure that promotes assessment literacy while raters score samples. Furthermore, using modern measurement theory, we strengthen the field's understanding of writing assessment by providing evidence of validity, reliability, and fairness of scores to support the interpretation and use of the WRITE.

Keywords: writing assessment, teacher education, preservice teachers, psychometric theory, reliability, validity, fairness

1. Introduction

Despite the importance of writing instruction and assessment in primary, secondary, and post-secondary education, very few teacher preparation programs include adequate instruction about writing instruction and assessment (Martin & Dismuke, 2018; Myers et al., 2016). While developing competent writers should be a goal in all higher education fields, it is particularly imperative for those working with preservice teachers. With only 25% of preservice teachers taking a course devoted to writing instruction (Myers et al., 2016), it may not be surprising that students in early elementary grades write less than 30 minutes per day (Coker et al., 2016). These findings reveal a concern in education – teachers may not be prepared to effectively assess and then teach writing, which impacts students' educational and professional success. Improving how teacher educators assess, and therefore develop, the writing skills of preservice teachers has the long-term potential of improving K-12 students' writing. To address this concern, the purpose of the present study is to validate a rubric for writing assessment, known as the Writing Rubric to Inform Teacher Educators (WRITE), that yields valid, reliable, and fair scores while simultaneously informing the rater about the qualities of effective writing.

Writing assessment is an under-researched challenge in teacher preparation programs (Ferris, 2014). When asked about forms of literacy assessment, educators are likely to discuss those related to reading, often omitting writing (Vallalón, Mateos, & Cuevas, 2015). If educators mention writing assessment, their intent is frequently either to assess English conventions or content knowledge related to another discipline (Guo, Crossley, & McNamara, 2013). Assessing actual writing skills is regularly reduced to examining grammar and spelling, with insufficient attention paid to organization, argument, or voice (Smith, Cheville, & Hillocks, 2006).

Previous studies indicate raters have a substantial impact on writing assessment outcomes (e.g., Janssen, Meier, & Trace, 2014). Rather than focusing solely on whether raters scored performances consistently, many writing assessment researchers are concerned with observing patterns related to construct-relevant and construct-irrelevant variables, which modern measurement theory refers to as "fairness" (American Educational Research Association [AERA], National Council on Measurement in Education [NCME], & American Psychological Association [APA], 2014). According to the *Standards for Educational and Psychological Testing* put forth by the AERA, APA, and NCME, "a test that is fair minimizes the construct-irrelevant variance associated with individual characteristics and testing contexts that would otherwise compromise the score validity for some individuals" (p. 219). Researchers acknowledge that rater judgment plays a mediating role in the assessment of students' writing proficiency (Author, 2018). In other words, many researchers and practitioners are concerned with the degree to which rater errors and systematic biases introduce construct-irrelevant variance into the interpretation of ratings—thus threatening the evidence of reliability, validity, and fairness of any writing assessment (Attali, 2016; Author, 2013; Trenary & Farrar, 2016). Psychometrically sound writing assessments facilitate *rater-invariant* interpretations of student achievement—that is, students' estimated writing achievement does not depend on the rater who scored their essay, and raters' estimated severity does not depend on the essays they scored.

Historically, researchers have focused their energy on how to improve rating quality, assuming that the problem resides in the fact that raters were not well-trained (Attali, 2016; Rezaei & Lovorn, 2010; Wolfe, Mathews, & Vickers, 2010). However, through multiple studies, the consensus remains that even through rigorous training and calibration methods, raters differ in severity (Attali, 2016; Eckes, 2008; Engelhard & Myford, 2003). A rater will even score a sample differently at different time points (Leckie & Baird, 2011; Myford & Wolfe, 2009; Wolfe, Moulder, & Myford, 2001). These findings indicate that focusing on rater training may not be enough to ensure psychometrically sound writing assessments. Other construct-irrelevant variables, such as context, may contribute to rater judgments (Attali, 2016). In addition to rater effects, it is also possible that characteristics of instruments (e.g., rubrics) can contribute to psychometric limitations in rater-mediated writing assessments.

2. Purpose

To address concerns with teaching writing to future teachers, the purpose of the present study is to validate a rubric for writing assessment, known as the Writing Rubric to Inform Teacher Educators (WRITE), that yields valid, reliable, and fair scores while simultaneously informing the rater about the qualities of effective writing. We developed a rubric that *informs* raters about high-quality writing throughout the scoring process—thus minimizing discrepancies among raters due to different interpretations of the instrument. Our rubric is unconventional in its design, and we hypothesize that, coupled with calibration and training, the Writing Rubric to Inform Teacher Educators (WRITE) contributes to psychometrically sound rater-mediated writing assessment systems across raters and samples. For the present study, we focus on the function of this newly-created rubric by showing that the WRITE shows evidence of reliability and validity while also adhering to evidence of fairness through meaningful differences across raters.

3. Literature Review

To better understand the current literature supporting our study, we provide a review of research looking at three lines of research inquiry: (1) assessment literacy in teacher education, (2) concerns with scoring procedures, and (3) evaluating existing writing rubrics to develop the WRITE. We also provide the models of writing that informed the inclusion and exclusion of specific constructs for the WRITE. In the following sections, we present our synthesis of the prior research on these topics that informed our development of the WRITE.

3.1. Assessment Literacy in Teacher Education

In teacher education, writing research has not emphasized assessment literacy to a degree that would help teacher educators and preservice teachers understand the technical and affective variables needed and implemented in scoring writing. Writing assessment researchers define assessment literacy as a complex intersection of "technical know-how, practical skills, theoretical knowledge, and understanding of principles" (Taylor, 2009, p. 27) with special attention to context. Therefore, writing assessment should be conducted by professionals with writing skill knowledge, an understanding of how to write, and contextualized knowledge about how the writing was developed and completed.

One component of assessment literacy identifies rater training and calibration as an important element in documenting consistent, useful scores from a rubric. In her early work, Weigle (1998) found that inexperienced raters (i.e., those with limited training and calibration) scored writing samples with more severity and less internal (intra-rater) consistency than those raters who were "normed" to the scoring process (i.e., experienced raters). After substantial training, the difference in scores (i.e., rater severity) between the two groups decreased, yet differences persisted.

While Weigle's (1998) work provided important insight into the effects of rater training on rater severity and intra-rater consistency, she did not emphasize how the differences in rater scores could be interpreted and useful. These considerations are related to the fairness of rater scores, which indicate that across raters, items, and test-takers (e.g., samples), scores are consistent.

In a 2016 special issue of the *Journal of Writing Assessment*, researchers Kelly-Riley and Whithaus (2016) emphasized the education field's minimal attention on fairness and over-reliance on reliability and validity evidence. Many of the concerns related to fairness center on whether instruments, or rubrics, provide all learners with ethical and fair accounts of their achievement (Elliot, 2016; Slomp, 2016). In assessment literacy, teacher educators must consider these ethical implications, but in terms of measurement alone, we also argue that modern measurement theory can psychometrically address concerns with scoring fairness. In other words, the rubric function must support diversity in scores while also acknowledging the scorers' expertise. The measurement implications of fairness in scoring have been over-looked in assessment literacy and writing assessment research.

Crusan, Plakans, and Gebril (2016) state that "assessment literacy is not just about content or delivery but how this content is enmeshed with teachers' knowledge, beliefs, and practices" (p. 45). Most rubrics, to date, assume that teachers or teacher educators have all the knowledge, experience, and practice to score writing effectively; however, as prior research has shown a disconnect between scorers' knowledge and consistency, that conclusion may not be supported. Research has also demonstrated that many undergraduate and graduate teacher education programs do not require specific courses on writing instruction or writing assessment (Weigle, 2007). Therefore, assessment literacy may be improved if rubrics provided guidance within their form, structure, and layout to help scorers learn about effective writing. Currently, to our knowledge, rubrics with this function have not been created or proposed.

Focusing on an understanding of the current limitations to assessment literacy development of preservice teachers and teacher educators, the WRITE proposes to provide stronger evidence of score reliability, validity, and fairness through its unique structure. The WRITE teaches assessment literacy as scorers use the rubric, which we hypothesize increases consistency across raters.

3.2. Concerns with Scoring Procedures

Historically, researchers and practitioners have explored several systems to measure writing, from curated portfolios to timed writing samples (Barrot, 2016). Currently, the preferred method of scoring writing comes from human scorers using a rubric (Author, 2018a). While researchers have worked to streamline the writing scoring process, many concerns still exist including consistency among raters, samples, and items and processes for using automated scoring systems (Humphry & McGrane, 2015; Jimenez, 2017; Author, 2018b).

Researchers' and practitioners' primary concern with scoring writing samples is related to the consistency with which raters score student compositions (Author, 2017). Researchers have created various rubrics and scoring measures, finding evidence of rater consistency (Jimenez, 2017; Nielsen, Abbott, Griffin, Lott, Raskind, & Berninger, 2016). However, to date, researchers agree that little consistency among raters exists, even through rigorous practices to improve reliable scoring (see Author, 2018a). For example, in a study by Attali (2016), scorers were provided multiple calibration and training sessions over time. The results indicate that undergoing training either once, or throughout the scoring process, did not improve the consistency with which raters scored student performances. Therefore, evidence of reliable scores across raters may not be the only variable researchers should examine when creating rubrics to measure writing achievement.

We argue that much of the research on writing assessment has omitted an important element: fairness. Fairness relates to the role of human judgment in writing assessment (Deane, 2013; Guo et al., 2013). Specifically, the E.T.S. Standards for Quality and Fairness (2014) define fairness as "the extent to which the inferences made on the basis of test scores are valid for different groups of test takers" (p. 19). Fairness is the construct related to ensuring that each test taker is treated equally in regards to the instrument and that the instrument does not unduly penalize or reward any group or individual. Additionally, fairness relates to the raters by ensuring that they are trained in a way that allows them to score samples with an unbiased lens toward groups or individuals.

In psychometrics, fairness relates to the amount of construct-irrelevant differences that exist. These differences are not representative of the instrument or construct being used, but stem from biases the raters may bring to the scoring process. If a rubric is used to score writing and differences are seen among raters, modern measurement theory can be used to determine if those differences exist among specific individuals or groups, or are specific to certain constructs that may be unclear when using the rubric. One reason for construct-irrelevant differences in rater judgments may be that rubrics provide insufficient guidance about the writing process and writing components. Using rubrics to score writing assumes that the raters are highly trained and familiar with the writing content. However, we know this is not always true (Attali, 2016; Knoch, 2011; Lim, 2011; O'Sullivan & Rignall, 2007). In fact, few raters receive specialized training about writing instruction and content (Gallavan, Bowles, & Young, 2007; Hall, 2016; Morgan 2010; Myers, 2016).

When differences among raters exist, it is important to determine why those differences are present as a method for empirically evaluating the fairness of an assessment procedure. Raters may bring varying levels of expertise and background, along with their training, that may cause them to consistently rate certain constructs more severely than others. If these differences do not discriminate across test-takers, evidence of fairness will still be present. In other words, rather than viewing raters as "scoring machines" from which consistent judgments are the ultimate goal, raters can also be viewed as "expert judges," from whom construct-relevant differences can be used to provide more accurate student achievement estimates. If raters exhibit other evidence of psychometric quality, modern measurement theory techniques can be used to adjust for within-rater severity differences to provide valid, reliable, and fair estimates of student achievement (Humphry & McGrane, 2015; Jimenez, 2017). However, much of the language assessment research to date has emphasized procedures for quantifying rater disagreement (Author, 2017), rather than distinguishing between construct-relevant and construct-irrelevant differences among raters' judgments.

Fairness is psychometrically tied to both validity and reliability. Evidence of validity and reliability is critical, but insufficient for psychometrically defensible writing assessments (AERA, APA, & NCME, 2014). That is—evidence that writing assessment procedures result in appropriate interpretations and uses of assessment results (validity), and evidence of reproducible and precise estimates of student achievement (reliability) are inadequate without evidence that assessment results are not unduly influenced by construct-irrelevant characteristics (fairness). Many researchers and practitioners are rightly concerned with the degree to which raters exhibit consistency in writing assessment. However, modern measurement theory argues that evidence of rater consistency is not satisfactory evidence of psychometric quality.

3.3. Evaluating Existing Writing Rubrics to Develop the WRITE

Researchers agree that rubrics are an important component in establishing psychometrically sound writing assessments; however, researchers also struggle to agree on appropriate methods for developing rubrics (Nauman, Stirling, & Borthwick, 2011; Saddler & Andrade, 2004; Timmerman, Strickland, Johnson, & Payne, 2011). We consulted several related tools designed to measure writing achievement in the development of the WRITE (see Table 1). Within our analysis of various rubrics, we identified concerns in their structure or elements that strongly informed the development of the WRITE.

Trenary and Farrar (2016) developed a rubric to evaluate student writing samples in a nursing program to assess students' proficiency in writing mechanics (e.g., spelling, grammar, punctuation, APA Format, APA style, academic writing, citations, and references). Evaluating writing mechanics is a necessary part of assessment, but an exclusive focus on this component threatens the construct validity of the instrument, because additional writing elements (e.g., clarity, organization, etc.), should also be assessed.

The 6+1 Traits Writing Rubric also informed WRITE creation. As a popular means of writing instruction and assessment in K-12 contexts, the 6+1 rubric measures ideas, organization, voice, word choice, sentence fluency, conventions, and presentation. The traits and approach are applicable to all writing types but omits important elements of synthesis and evidence that are essential to students' growth as writers within discipline-specific contexts.

The Rubric for Scientific Writing (RSW; Author, 2016) proved most useful in providing an initial structure for our rubric creation. While focused specifically on K-12 science writing, the RSW assesses claim, evidence and support, analysis of content, organization, audience, and presentation of writing. The RSW provides a more holistic approach to writing assessment applicable to university level writing, while also aligning to the Common Core State Standards and Next Generation Science Standards. This rubric, therefore, combines standards-based instruction with more conventional elements of writing achievement.

To develop the WRITE, we utilized the structure of the RSW (Author, 2016) and analyzed limitations of prior rubrics, paying close attention to elements not emphasized and criteria overlooked. One similarity found among most previously published rubrics is that they assume expertise from the raters. No previously published rubric presented writing in a manner that taught the raters as they analyzed writing. We developed the WRITE in response to this limitation by presenting a method to teach elements of high-quality writing to the raters to emphasize traditional writing elements (e.g., conventions, organization) along with more advanced writing skills (e.g., synthesis, evidence, and grounding in the discipline).

3.4. Construct Model of Writing for Developing the WRITE

In addition to consulting several rubrics on writing, we also identified relevant theories supporting the constructs and models of writing in the WRITE. Specifically, we consulted two theories that promote writing achievement and development: (a) the Cognitive Processes Theory of Writing, and (b) the Sociocultural Theory of Writing. In the following paragraphs, we outline these two theoretical approaches, how they inform the WRITE, and break down the research regarding the constructs measured in the WRITE.

The Cognitive Processes Theory of Writing (Hayes, 1996, 2006; Flower & Hayes, 1981) describes a hierarchical thinking process that promotes and creates writing. This model emphasizes that writing must be taught in a manner that builds on previously learned skills, develop based on goal-setting, and incorporate both macro and micro levels of writing. Macro levels allow writers to consider the whole when writing, while micro levels allow writers to look at more specific details such as paragraph organization or word choice. In the WRITE, macro and micro levels of organization (that is, the purpose, synthesis, and presentation of writing) were all evaluated based upon the Cognitive Processes Model of Writing. By scoring strongly in these areas, a writer will display a high level of cognitive support for writing achievement.

The second theory, the Sociocultural Theory of Writing (Prior, 2006), states that effective writing is driven by cultural and societal norms and may differ across populations. This model of writing also stipulates that writers must be engaged and knowledgeable about the topic, and when learning to write, will learn from engaging in collaboration, building motivation, and seeing effective models. These principles led to the development of the *grounding in the discipline* and *evidence and support* sections of the WRITE. These two constructs may differ across populations, but through a U.S. lens, allow writers to support their thinking and show their background knowledge for a topic.

Finally, in addition to the two models described, the WRITE was influenced by the rubrics outlined in section 3.3. The 6+1 Traits (Coe, 2011) has long stood as an effective rubric for measuring K-12 students' writing, so we analyzed each of the traits to determine what would be most appropriate for preservice teacher level writing. From this rubric, we included the constructs of purpose, organization, and presentation of writing. We were also influenced by the work of Author (2016) and the Rubric for Scientific Writing. As this rubric emphasized writing for scientific and professional audiences, we also included the constructs of *evidence, synthesis*, and *grounding in the discipline* to allow writers to be members of the professional community in which they belong.

3.5. Developing a New Rubric to Fill the Gap in Research

Writing assessment is often critiqued for being subjective, due to the role of human judgments. Furthermore, answering the question "what is good writing" can be difficult and varies depending on who is asked (Nauman, Stirling, & Borthwick, 2011). However, writing subjectivity is not a reason for concern, if the raters disagree in construct-relevant ways that do not discriminate across groups or individuals. For instance, Author (2016) found that science teachers and English teachers, using the same rubric, scored the same piece of writing differently because they viewed the writing through their content area lens. For example, for one writing sample, science teachers emphasized figures the students drew as part of their written explanations and focused on whether the drawings and written explanations aligned. In contrast, English teachers emphasized the organization of ideas, coherence of the explanations, and synthesis of evidence for the argument; however, many English teachers did not focus heavily on the drawn depictions. While differences in scores existed due to these foci, both sets of teachers viewed the writing as experts in their fields. As unique content area experts, English teachers and science teachers could provide students with specific, meaningful feedback on their writing resulting in student growth, achievement, and possibly motivation to write (see Author, 2016). Through modern measurement theory, we identify how severe or lenient individual raters score writing samples and distinguish differences across raters, samples, and rubric items.

4. Methods

In the following sections, we describe the development of the WRITE. Then, we describe our sample, rating procedures, and data analysis procedures that show evidence of validity, reliability, and fairness across scores from the WRITE. Specifically, we sought to answer the following research questions:

1) What evidence is available to support the interpretation and use of ratings from the WRITE?
    a. To what extent is there evidence of validity to support the interpretation and use of ratings from the WRITE?
    b. To what extent is there evidence of reliability to support the interpretation and use of ratings from the WRITE?
    c. To what extent is there evidence of fairness to support the interpretation and use of ratings from the WRITE?

4.1. Developing the Writing Rubric to Inform Teacher Educators (WRITE)

The WRITE includes six elements (defined below) including: (1) purpose of the text; (2) level of grounding in the discipline; (3) organization; (4) evidence and support; (5) synthesis; and (6) presentation of writing (see Appendix A). Each element divides further into at least two sub-elements to provide deeper clarity, scoring, and assessment.

We developed these six elements by exploring prior research in writing and existing rubrics (see Table 1). To ensure our rubric captured components frequently missing in typical rubrics and aligned with the outcomes of K-12 writing, we utilized the Common Core State Standards (CCSS; National Governors Association, 2010), specifically focusing on the Anchor Standards for Writing. The CCSS Anchor Standards for Writing focus on opinion, narrative, persuasive, and argumentative writing. We added two elements to emphasize research-based and discipline-specific writing: synthesis and level of grounding in the discipline. From the standards and prior research, we also developed elements for purpose, organization, evidence and support, and presentation of writing (see Appendix A). We limited our mastery levels to three categories: (1) mastery, (2) developing, and (3) novice.

4.2. Rating Procedures

3.2.1. Participants. We recruited, trained, and calibrated four expert raters to use the WRITE. We define expert raters as individuals with substantial knowledge of assessment literacy, writing instruction, and writing assessment. Our goal with these raters was to determine if scores on the WRITE would show meaningful patterns and trends across raters, indicating evidence of fairness in the scores. The four raters were literacy teacher educators with former experience as elementary or middle school teachers, who currently teach literacy courses in preservice teacher preparation programs in the United States. Therefore, all four raters were experts both in the field of literacy and with assessing preservice teacher writing.

4.2.2. Writing Samples. We collected two writing samples each from 23 preservice teachers enrolled in an upper-level literacy course at a mid-sized institution in the Southeastern United States. At least three raters scored each of the blinded 46 samples. The writing samples represented essays written for a general literacy methods course within a teacher preparation program and were each three to five pages in length. We did not designate how these writing samples should be composed, as we wanted to collect authentic examples of preservice teacher writing.

4.2.3. Evidence of Reliability. To maximize reliability, the lead researcher created professional development materials to aid in training and calibration for using the WRITE. These materials included a WRITE Handbook, outlining scoring procedures, and helpful tips for evaluating student writing. The WRITE Handbook provides a step-by-step guide for determining the level of mastery for each element and provides examples of questions and considerations to aid users of the WRITE in scoring samples. We distributed this handbook to all raters prior to scoring.

To validate the WRITE during the development process, a team of literacy faculty conducted several rounds of calibrations. First, each rater practiced scoring samples of de-identified student work using the WRITE. These samples were not part of the 46 samples in the present study but were pulled from a comparable course. The research team independently scored the samples then met to discuss limitations and strengths of the WRITE. This process was repeated three times over one semester to allow the team to understand WRITE scoring procedures. Once the raters reached over 90% agreement on a set of scores, which occurred during the third round of scoring, raters scored a class set of writing samples to determine inter-rater reliability. Overall, between four raters, inter-rater reliability ranged from 0.833 to 0.953 agreement, depending on the element (see Table 2). Overall, the average inter-rater reliability across the entire WRITE was 0.891. Even with formerly untested writing elements such as grounding in the discipline, the WRITE shows evidence of reliable scores among raters.

4.2.4. Scoring Procedures. After reaching consensus and establishing acceptable inter-rater reliability, the four raters independently scored 46 writing samples. Each sample was scored by three of the four raters to determine severity among raters, distribution of scores on individual components of the WRITE, and overall score differences. The raters scored samples in bundles of six, and then met to discuss their scores and justifications. These discussions allowed raters to better understand scoring procedures and the WRITE components. At each meeting, the raters were provided with additional training and calibration if the scores varied or if the justifications deviated from the WRITE. The goal of the meetings and calibrations was to inform and teach the raters about high-quality writing and score usage, not to influence their judgments as expert raters.

4.3. Analytic Approach

We used a measurement model based on Rasch measurement theory (Rasch, 1960) to evaluate the psychometric properties of the WRITE instrument. We used this approach for several reasons. First, researchers and practitioners frequently use models based on Rasch measurement theory to evaluate the quality of rater-mediated writing assessments (e.g., Author, 2018a; Myford & Wolfe, 2003; Wolfe, Song, & Jiao, 2016). Second, Rasch models allow researchers to investigate the psychometric properties of rater-mediated writing assessments using evidence beyond group-level indicators of rater reliability. Finally, to promote fairness, it is essential that differences in rater severity be accounted for when estimating student achievement. Methods based on latent variable models in general, and Rasch models in particular, provide such evidence and facilitate these comparisons.

We used a three-facet formulation of the Many-Facet Rasch (MFR) model (Linacre, 1989) in our analysis:

$$\ln\left[\frac{P_{nji(x=k)}}{P_{nji(x=k-1)}}\right]=\theta_n-\lambda_j-\delta_i-\tau_{ik}, \quad (1)$$

where the left side of the equation indicates the log of the odds that student *n* receives a rating in rating scale category *k* ($x = k$), rather than in the category just below it ($x = k – 1$) from rater *j* on item i, θ_n is the estimated location (i.e., judged writing achievement) of test-taker *n*, λ_j is the estimated location (i.e., severity) of rater *j*, and δ_i is the estimated location (i.e., judged item difficulty) of item *i.* The final term in Equation 1 (τ_{ik}) is the estimated logit-scale location at which the probability for a rating in category *k* is equal to the probability for a rating in category $k – 1$, specific to item *i.* For each item, we estimated two thresholds: τ_1 is the threshold between the first and second rating scale categories, and τ_2 is the threshold between the second and third rating scale categories.

It is important to note that we specified the MFR model in Equation 1 such that the rating scale element thresholds (τ) are estimated separately for each of the items in the WRITE rubric. We used this approach so that we could empirically evaluate the degree to which the rating scale had a comparable structure over the individual items in the rubric. Rather than assuming the raters' interpretation of category difficulty was similar for each of the items, this approach allowed us to examine how the rating scale functioned empirically. We used the Facets software program (Linacre, 2015) to analyze our data according to the MFR model in Equation 1.

4.3.1. Indicators of Psychometric Quality. We focused on three categories of indicators of psychometric quality: (1) locations and precision, (2) model-data fit of the location estimates, and (3) rating scale category functioning.

4.3.1.1. Locations and Precision. First, we examined the estimated locations for test-takers (achievement), raters (severity), and items (difficulty), along with the precision of these estimates. In our study, high locations on the logit scale indicate high levels of judged writing achievement, more-severe raters, and more-difficult items.

Relatedly, we examined the precision of the estimates for each facet in Equation 1. We used three indicators of precision: (1) targeting, (2) standard errors (*SE*), and (3) reliability of separation statistics (*Rel*). First, targeting refers to the alignment between the distributions of test-taker, rater, and item estimates. Evidence of close alignment between these distributions provides support for the interpretation of the location estimates. Second, *SE*s are numeric indicators of precision for each element within each of the facets. Small values of the *SE* indicate that a particular element within a facet (e.g., an individual test-taker) has been measured with enough precision to confidently interpret their logit-scale location. Finally, we calculated reliability of separation (*Rel*) statistics as an indicator of the degree to which there were substantively meaningful differences in the location estimates of individual test-takers, raters, and items. When there is acceptable model-data fit (described further below), *Rel* for test-takers is comparable to coefficient alpha. For raters and items, *Rel* describes the degree to which individual raters and individual items have distinct locations. In rater-mediated writing assessments, high values of *Rel* indicate that there is a meaningful order of test-takers, raters, and items on the continuum that reflects the construct. As long as there is adequate model-data fit for raters, test-taker achievement estimates are adjusted for differences in rater severity—such that the interpretation of student achievement does not depend on the particular rater who rated their essay. In other words, the estimation procedure for calculating student location estimates controls for the severity of the rater(s) who scored a student's performance (Linacre, 1989). This rater-invariant interpretation of student achievement is an essential component of the fairness of rater-mediated writing assessments.

4.3.1.2. Model-Data Fit. In general, researchers who use Rasch models evaluate *model-data fit* using summaries of residuals, or differences between observed ratings and the ratings that would be expected using the logit-scale location estimates. The most popular approach to summarizing model-data fit for Rasch models is the use of mean square error (*MSE*) statistics, including Infit *MSE* and Outfit *MSE*. Because Outfit *MSE* statistics are relatively more sensitive to extreme unexpected ratings, we focus on this form of *MSE* statistics in this study. For raters, Outfit *MSE* is calculated as follows:

$$\text{Outfit } MSE = \frac{\sum_{n}^{N} Z_{nji}^{2}}{N} \quad , \qquad (3)$$

where Z_{nji} is the standardized residual (difference between observed and expected rating) for Rater *j*'s rating of student *n* on item *i*, and *N* is the number of test-takers. One can calculate Outfit *MSE* for test-takers and items using the sum of the residuals associated with a particular test-taker or item, respectively.

Researchers and practitioners who use Rasch models generally agree that values of Outfit *MSE* around 1.00 are expected when there is adequate model-data fit. Further, for rating scale data, values between 0.5 and 1.5 are generally considered evidence of acceptable model-data fit (Author, 2018a). In this study, we treat Outfit *MSE* as a continuous variable, while recognizing these recommendations.

4.3.1.3. Rating Scale Functioning. The last category of evidence of psychometric quality is rating scale functioning. Evidence in this category helped us to empirically evaluate the degree to which the WRITE rubric rating scale categories functioned comparably over the items. First, we compared the location estimates of the rating scale category thresholds (τ_{ik}) across items. This analysis allowed us to evaluate whether the difference in difficulty between rating scale categories (e.g., the difference in a rating of *Novice* and *Proficient/Developing*) had a similar interpretation over the WRITE rubric items. Second, we examined the locations of the rating scale category thresholds for evidence that the rating scale categories were ordered as expected. Finally, we examined the locations of the rating scale category thresholds for evidence that the raters used each category to describe a distinct range of writing achievement.

5. Results

Overall, the data displayed acceptable fit to the MFR model. Specifically, the logit-scale location estimates explained 39.19% of the variance in raters' ratings; this value is well above the critical value of 20% that Reckase (1979) suggested for Rasch analyses of potentially multidimensional scales.

5.1 Locations and Precision

Table 3 includes a summary of the logit-scale locations and *SE*s for students, raters, and items. On average, the student locations were close to the rater and item locations (all around zero logits) on the logit scale—suggesting acceptable targeting between the three facets. One can glean additional information about logit-scale locations using Figure 1. Specifically, Figure 1 is a *variable map*, which is a graphical display of the logit-scale locations of individual test-takers, raters, items, and rating scale category thresholds. The first column shows the logit scale on which the test-takers, raters, and item locations have been estimated. Higher numbers indicate higher judged writing achievement, more-severe raters, more-difficult items, and more difficult rating scale categories.

The second column in Figure 1 shows the locations for individual test-takers, where each test-taker is represented using an asterisk. Examination of the test-taker locations reveals a wide range of locations, between -2.94 logits for the test-taker with the lowest judged writing achievement (mean rating = 1.26) and 3.23 logits for the test-taker with the highest judged writing achievement (mean rating = 2.79). The third column shows the locations for the individual raters. The location estimates reflect differences in rater severity, ranging from -0.47 logits for the most-lenient rater (Rater 3; mean rating = 2.13) to 0.34 logits for the most-severe rater (Rater 4; mean rating = 1.87). The fourth column shows the locations for the individual items. Examination of these estimates reveals a range of item difficulties, between -1.32 logits for the item that was judged as easiest (Item 6b; mean rating = 2.33) and 1.56 logits for item that was judged as most difficult (Item 5a; mean rating = 1.54). The final 12 columns in the variable map illustrate the calibration of the rating scale categories. We describe these columns later in the Results section.

To evaluate the precision of the location estimates, we used *SE*s and separation statistics. Returning to Table 3, we observed small *SE*s for test-takers, raters, and items, relative to the spread of logit-scale locations in our analysis. Furthermore, the *SE*s for all three facets were within the range that researchers have reported in previous analyses of rater-mediated assessments with incomplete rating designs (e.g., Author, in press). Specifically, the average *SE* for the student facet was relatively larger for than the average *SE*s for the rater and item facets. This result is to be expected, given that there were more observations of each rater and item compared to each student. Finally, the relatively high reliability of separation statistics for the three facets indicated that there were meaningful differences in student achievement (*Rel* = 0.94), rater severity (*Rel* = 0.86), and item difficulty (*Rel* = 0.91). With regard to the relatively high *Rel* statistic for raters, two observations are important to note. First, the spread of the raters' locations on the logit scale overlaps considerably with the spread of student locations. Although there were differences between the individual raters in terms of the severity with which they evaluated student performances, the targeting between students and raters facilitated precise measures of student achievement (Embretson & Reise, 2000). Second, the differences among raters are controlled during the estimation procedure for the Rasch model—thus minimizing the effect of very severe or very lenient raters on conclusions about individual student achievement (Linacre, 1989).

5.2. Model-Data Fit

Table 3 provides a summary of the Outfit *MSE* statistics for each facet. For students, the average Outfit *MSE* statistics were close to the generally accepted value of 1.00—indicating acceptable overall model-data fit for students. Further examination of Outfit *MSE* statistics for individual students revealed that these statistics ranged from 0.55 for the student whose essay ratings were most consistent with model expectations to 1.84 for the student whose essay ratings deviated most from model expectations.

For raters, the average Outfit *MSE* statistic was also around 1.00—indicating acceptable global fit for this facet. For the individual raters, these statistics ranged from 0.79 for the rater whose ratings were somewhat more consistent than expected by the model to 1.34 for the rater whose ratings were most deviant from model expectations. Nonetheless, fit statistics for individual raters were within the range of critical values that several researchers have proposed for evaluating rater fit in performance assessments (e.g., Author, 2018a). The important implication of this evidence of acceptable model-data fit for the rater facet is that, although there were notable differences in rater severity, the estimates of student achievement were meaningfully adjusted to account for these differences as part of the Rasch model estimation procedure (Linacre, 1989).

We also observed acceptable model-data fit for items (around 1.00). For individual items, Outfit *MSE* ranged between 0.88 to 1.31—such that all items had model-data fit statistics within the generally accepted range for rating scale data.

5.3. Rating Scale Functioning

As an initial step in evaluating rating scale functioning for the WRITE, we examined the frequency with which the raters used the three rating scale categories. Across the items, 22% of the ratings were in Category 1, 57% of the ratings were in Category 2, and 21% of the ratings were in Category 3. With this evidence, we proceeded to examine the degree to which the WRITE rubric rating scale categories functioned comparably over the individual items. First, we examined the locations of the rating scale category thresholds on the logit scale across all the items. The last 12 columns of Figure 1 show these locations—where each column shows the rating scale structure for one item (e.g., "I.1a" shows the rating scale structure for Item 1a). In each of the rating scale columns, horizontal lines show the location of the threshold between adjacent rating scale categories. Although the items are expected to have different levels of difficulty, a consistent interpretation of the *distance between* rating scale categories across items depends on consistent locations of these thresholds. Examination of the threshold locations in Figure 1 indicates that, in general, the rating scale category thresholds are comparable across the items, because the horizontal lines share similar locations. However, there are some items for which the category locations are slightly different. For example, the middle rating scale category is somewhat wider for item 2b compared to item 2a—indicating that a rating in the second category relative to the first category is easier, and a rating in the third category relative to the second category is somewhat more difficult for item 2b compared to item 2a.

Table 4 provides additional information about rating scale category functioning. Specifically, this table shows that the raters used all three rating scale categories when they rated student compositions. Table 4 also provides evidence about the directionality of the rating scale categories. First, the table shows the average logit-scale location for test-takers who received a rating in each rating scale category for each item. Second, the table shows the estimated threshold locations for each item, where τ_1 is the threshold between the first two rating scale categories, and τ_2 is the threshold between the second and third rating scale categories. The results in Table 4 indicate that the WRITE rating scale categories are functioning as expected because the average logit-scale locations increase as rating scale categories increase, and τ_1 is lower than τ_2 for all 12 items. Finally, Table 4 shows the absolute value of the distance between the two rating scale category thresholds for each item ($|\tau_1 - \tau_2|$). These distances provide information about the degree to which the rating scale categories describe a distinct range of writing achievement. All of the distances are between 1.4 logits and 5 logits, which is the generally accepted range of effective distances between rating scale category thresholds (Linacre, 2002).

6. Discussion

Researchers who have evaluated writing assessment rubrics have emphasized evidence of reliability and validity and have focused less on issues of fairness related to rater-mediated assessments. The WRITE shows strong evidence of validity, reliability, and fairness of scores, which we discuss further in the next sections.

6.1. Evidence of Validity

Through stages of development, we focused on creating a rubric that emphasized assessment literacy while showing strong evidence of consistent scores among raters. To accomplish this goal for the WRITE, we determined that the rubric showed strong evidence of validity in the interpretation of scores. Our process included multiple steps. First, we consulted prior rubrics used in research (see Table 1) and created a comprehensive database of elements typically assessed by rubrics, as well as those omitted. Second, for alignment with K-12 education, we consulted the CCSS Anchor Standards for Writing, prior research on writing, and the National Writing Project (NWP). Third, once the WRITE was created, we had the rubric reviewed by literacy professionals to verify that it represented the current discussions on writing research, theory, and practice. Fourth, for each element, we outlined a set of general criteria to assist in increasing consistency of scoring and aid both teachers and students in recognizing mastery attributes of each element. Finally, through the present study, we noted how the raters were able to use the WRITE effectively, efficiently, and consistently.

6.2. Evidence of Reliability

Evidence of reliability is essential for a rubric to be useful to teachers, students, and scholars. Previous writing rubrics such as the 6+1 Traits (Coe et al., 2011) and the ACT scoring rubric (ACT Writing Test, 2009) have shown strong evidence of reliability. For example, the 6+1 Traits (Coe et al., 2011) rubric has an interrater reliability of 97% while the ACT scoring rubric has a Cronbach's alpha score of 0.91. Both indicate strong evidence of reliability. Scores from the WRITE yielded a Cronbach's alpha reliability estimate of 0.891. With thirteen total items, including previously untested items, this score shows moderate evidence of reliability.

Within our study, we noted that reliability coefficients may not provide all the evidence needed to make decisions about the effectiveness of a writing rubric. Therefore, we also examined how meaningful the differences were among elements and across samples. Using a MFR model, we observed strong evidence of reliability using indicators of precision, targeting, and separation of estimates on the latent variable. For example, across elements, scorers consistently rated the synthesis element most severely. Across samples, there were a few samples that raters consistently scored higher or lower, again showing meaningful differences in individual students' writing which show achievement differences.

Therefore, our analysis indicates that the WRITE has strong evidence of reliability, using modern measurement theory as our rationale. Prior rubrics often only report interrater reliability or reliability coefficients, but do not examine if any meaningful differences exist across elements or samples.

6.3. Evidence of Fairness

Writing rubrics should show evidence of validity and reliability, but psychometric fairness across independent raters must match these scores. As we mentioned previously, many rubrics do not consider evidence of fairness, or how independent raters utilize the rubric (Hawthorne, Bol, & Pribesh, 2017). In other words, previously published work on writing achievement ignores the patterns in which raters may differ in their scoring. As we argue, those disagreements could show meaningful interpretations of the rubric.

Overall, we found that raters did differ in their level of severity. However, model-data fit statistics for the raters indicated that our raters were consistent in those levels, meaning a rater that was more lenient for one sample or WRITE element tended to always be more lenient. This finding is important because it shows that while raters interpreted the scoring of samples slightly differently, they were consistent. This consistency was maintained even over time, as the raters scored samples in bundles of six over the course of several months. As a result, the estimates of preservice teacher writing achievement were meaningfully adjusted to account for differences in rater severity—thus facilitating fair comparisons among preservice teachers whose writing samples were rated by different raters.

We also evaluated the degree to which the WRITE rating scale functioned effectively. The WRITE includes three rating scale categories. We observed empirically that the categories had a generally comparable structure across the items, that the categories reflected increasing levels of achievement, and that the raters were able to use the categories effectively to distinguish among preservice teachers with different levels of writing achievement. Together, the results from our analyses, indicate that the WRITE shows strong evidence of fairness across raters, test-takers, and items.

These results indicate that the WRITE provides consistent scores across raters while acknowledging meaningful differences across scores for each measured element. In other words, the WRITE is effective at distinguishing among test-takers, but there are different levels of rater severity. Our results indicate that the experiences and knowledge of the raters will influence how they score writing samples. Even though the individual raters exhibited different levels of severity, the raters still provided meaningful judgments of test-takers' writing achievement. Rather than relying on an automated scoring system with which raters do not engage with the instrument, our results show the value in human raters. While their levels of severity are different, the raters in our sample exhibited consistent levels of severity over different test-takers (good model-data fit), and they were consistent in their interpretation of the rating scale categories across items—thus providing initial evidence of fairness for the WRITE. Continued use of the WRITE could provide more evidence that the WRITE is a useful tool that is fair across groups and individuals.

Just as an assessment instrument is never "valid" or "invalid" (AERA et al., 2014), an assessment cannot be proven "fair" or "unfair." Rather, it is essential that test developers and test users gather empirical evidence to support the fairness of the interpretation of test scores for each intended use. In this study, we used a Rasch measurement approach to calibrate the WRITE and gather evidence of psychometric quality related to validity, reliability, and fairness. We observed that analyzing data the WRITE using a modern measurement approach that accounts for differences in rater severity resulted in desirable psychometric properties that support fairness. However, any interpretation and use of this instrument warrants empirical examination of evidence that is appropriate to the intended interpretation and use, specific to the population of test-takers who are of interest.

6.4. Limitations and Future Research

Some limitations exist in the present study. First, further evidence and support in multiple contexts is needed to determine the generalizability of the psychometric properties of the WRITE. Future research can provide additional evidence that the scores from the WRITE show evidence of being valid, reliable, and fair. Specifically, future research can provide evidence of validity by showing that the WRITE captures writing elements omitted from other rubrics. For example, researchers may score the same writing samples using the WRITE and additional rubrics to determine how effectively, efficiently, and consistently raters score with each rubric.

With regard to reliability, validity, and fairness, the raters in our study exhibited different levels of severity when they used the WRITE to score students' compositions. From a statistical perspective, the Rasch model mitigated the differences in rater severity as part of the estimation procedure for calculating student achievement (Linacre, 1989). However, additional research is warranted to glean substantive explanations for differences in the severity with which raters apply the WRITE in order to determine the extent to which severity differences reflect construct-relevant or construct-irrelevant factors. For example, in a future study, researchers could conduct think-aloud interviews with raters regarding their judgmental processes in order to understand how raters interpret and apply the WRITE to student compositions. Additionally, future research is needed across populations to determine if the WRITE exhibits fairness across groups and individuals.

7. Conclusion

Writing rubrics have often over-estimated the level of assessment literacy of scorers, which the WRITE specifically addresses. With its design focused on questions and teaching elements of effective writing, the WRITE teaches assessment literacy pertaining to writing assessment as raters score. Moreover, through calibration and training, we found that expert raters, with content knowledge of writing showed strong evidence of validity, reliability, and fairness across their scores. With future research, we hope to extend these results to teacher educators with less specific training and knowledge of writing instruction to further demonstrate the useful and applicability of such a unique rubric.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Attali, Y. (2016). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing, 33*(1), 99-115. doi: 10.1177/0265532215582283

Author. (2013). [information intentionally removed for peer-review]

Author. (2016). [information intentionally removed for peer-review]

Author. (2017). [information intentionally removed for peer-review]

Author. (2018a). [information intentionally removed for peer-review]

Author. (2018b). [information intentionally removed for peer-review]

Barrot, J. S. (2016). Using Facebook-based e-portfolio in ESL writing classrooms: Impact and challenges. Language, Culture, and Curriculum, 29(3), 286-301. doi: 10.1080/07908318.2016.1143481

Beach, R., & Friedrich, T. (2006). Response to writing. In C. A. MacArthur, S. Graham, & J. Fitzgerald's (Eds.) *Handbook of writing research* (pp. 222-234). New York, NY: The Guilford Press.

Coe, M., Hanita, M., Nishioka, V., & Smiley, R. (2011). An Investigation of the Impact of the 6+1 Trait Writing Model on Grade 5 Student Writing Achievement. Final Report. NCEE 2012-4010. *National Center for Education Evaluation and Regional Assistance*.

Coker, Jr., D. L., Farley-Ripple, E., Jackson, A. F., Wen, H., MacArthur, C. A., & Jennings, A. S. (2016). Writing instruction in first grade: An observational study. *Reading & Writing, 29*, 793-832. Doi: 10.1007/s11145-015-0506-6

Crusan, D., Plakans, L., & Gebril, A. (2016). Writing assessment literacy: Surveying second language teachers' knowledge, beliefs, and practices. *Assessing Writing, 28,* 43-56. doi: 10.1016/j.asw.2016.03.001

Cutler, L., & Graham, S. (2008). Primary grade writing instruction: A national survey. *Journal of Educational Psychology*, *100*(4), 907-919.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing, 18*(1), 7-24.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*(2), 155-185.

Elliot, N. (2016). A theory of ethics for writing assessment. *The Journal of Writing Assessment, 9*(1).

Engelhard Jr, G., & Behizadeh, N. (2012). Exploring the alignment of writing self-efficacy with writing achievement using rasch measurement theory and qualitative methods. *Journal of applied measurement, 13*(2), 132-145.

Engelhard, G., & Myford, C. M. (2003). Monitoring faculty consultant performance in the advanced placement English Literature and composition program with a many-faceted Rasch model. ETS Research Report Series, 2003(1).

Ferris, D. R. (2014). Responding to student writing: Teachers' philosophies and practices. *Assessing Writing, 19*(1), 6-23.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication, 32*(4), 365-387.

Gallavan, N. P., Bowles, F. A., & Young, C. T. (2007). Learning to write and writing to learn: Insights from teacher candidates. *Action in Teacher Education, 29*(2), 61-69.

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing, 18*(3), 218-238.

Hall, A. H. (2016). Examining shifts in preservice teachers' beliefs and attitudes toward writing instruction. *Journal of Early Childhood Teacher Education, 37*(2), 142-156.

Hawthorne, K. A., Bol, L. & Pribesh, S. (2017). Can providing rubrics for writing tasks improve developing writers' calibration accuracy?, *The Journal of Experimental Education, 85*(4), 689-708, doi: 10.1080/00220973.2017.1299081

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. M. Levy, & S. Ransdell (Eds.). *The science of writing: Theories, methods, individual differences, and applications,* (pp. 1-27). Mahwah, NJ: Erlbaum.

Hayes, J. R. (2006). New directions in writing theory. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.). *Handbook of writing research,* (pp. 28-40). New York: Guilford.

Humphry, S. M., & McGrane, J. A. (2015). Equating a large-scale writing assessment using pairwise comparisons of performances. *Australian Association for Research in Education, 42*, 443-460. doi: 10.1007/s13384-014-0168-6

Janseen, G., Meier, V., & Trace, J. (2014). Classical test theory and item response theory: Two understandings of one high-stakes performance exam. *Colombian Applied Linguistics Journal, 16*(2), 167-184.

Jimenez, J. E. (2017). Early grade writing assessment: An instrument model. *Journal of Learning Disabilities, 50*(5), 491-503. doi: 10.1177/0022219416633127

Kelly-Riley, D., & Whithaus, C. (2016). Introduction to a special issue on a theory of ethics for writing assessment. *The Journal of Writing Assessment, 9(*1).

Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior – a longitudinal study. *Language Testing, 28*(2), 179–200.

Koutsoftas, A. D., & Gray, S. (2012). Comparison of narrative and expository writing in students with and without language-learning disabilities. *Language, Speech, and Hearing Services in Schools, 43* 395-409.

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing, 28*, 543– 560.

Martin, S. D., & Dismuke, S. (2018). Investigating differences in teacher practices through a complexity theory lens: The influence of teacher education. *Journal of Teacher Education, 69*(1), 22-39. doi: 10.1177/0022487117702573

Morgan, D. N. (2010). Preservice teachers as writers. *Literacy Research and Instruction, 49*(4), 352-365.

Myers, J., Scales, R. Q., Grisham, D. L., Wolsey, T. D., Dismuke, S., Smetana, L., Yoder, K. K., Ikpeze, C., Ganske, K., & Martin, S. (2016). What about writing? A national exploratory study of writing instruction in teacher preparation programs. *Literacy Research and Instruction, 55*(4), 309-330. Doi: 10.1080/19388071.2016.1198442

National Center for Education Statistics. (2012). The nation's report card: Writing 2011. Washington, DC: U.S. Department of Education, Institute of Educational Sciences. National Commission on Writing. (2004). Writing and school reform. Retrieved from http://www.collegeboard.com

National Commission on Writing. (2005). Writing: A powerful message from state government. Retrieved from http://www.collegeboard.com

National Governors Association. (2010). *Common core state standards for English language arts & literacy in history/social studies, science, & technical studies.* Washington, DC: National Governors Association Center for Best Practices, Council of Chief State School.

Nauman, A. D., Stirling, T., & Borthwick, A. (2011). What makes writing good? An essential question for teachers. *The Reading Teacher, 64*(5), 318-328. doi: 10.1598/RT.64.5.2

Nielsen, K. (2015). Teaching Writing in Adult Literacy: Practices to Foster Motivation and Persistence and Improve Learning Outcomes. *Adult Learning, 26*(4), 143-150.

Nielsen, K., Abbott, R., Griffin, W., Lott, J., Raskind, W., & Berninger, V. W. (2016). Evidence-based reading and writing assessment for dyslexia in adolescents and young adults. *Learning Disabilities: A Multicultural Journal*, *21*(1), 38-56. doi: 10.18666/LDMJ-2016-V21-I1-6971.

O'Sullivan, B., & Rignall, M. (2007). Assessing the value of bias analysis feedback to raters for the IELTS writing module. In L. Taylor & P. Falvey (Eds.), IELTS Collected Papers (pp. 446–476). Cambridge: Cambridge University Press.

Penner, I. S. (2013). Comparison of effects of cognitive level and quality writing assessment (CLAQWA) rubric on freshman college student writing. *College Student Journal*, 447-461.

Prior, P. (2006). A sociocultural theory of writing. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.). *Handbook of writing research*, (pp. 54-66). New York: Guilford.

Reckase, M. D. (1979). Unifactor Latent trait models applied to multifactor tests: Results and implications. *Journal of Educational and Behavioral Statistics*, *4*(3), 207-230.

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing writing*, *15*(1), 18-39.

Ryan, M. (2014). Reflexive writing: Re-thinking writing development and assessment in schools. *Assessing Writing, 22,* 60-74.

Saddler, B., & Andrade, H. (2004). The writing rubric: Instructional rubrics can help students become self-regulated writers. *Educational Leadership,* 48-52.

Slomp, D. (2016). Ethical considerations and writing assessment. *The Journal of Writing Assessment, 9*(1).

Smith, M. W., Cheville, J., & Hillocks, G. (2006). "I guess I'd better watch my English": Grammar sand the teaching of the English language arts. In C. A. MacArthur, S. Graham, & J. Fitzgerald's (Eds.) *Handbook of writing research* (pp. 263-274). New York, NY: The Guilford Press.

Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics*, *29*, 21-36. doi: 10.1017/S026719050909003

Timmerman, B. E. C., Strickland, D. C., Johnson, R. L., & Payne, J. R. (2011). Development of a "universal" rubric for assessing undergraduates' scientific reasoning skills using scientific writing. *Assessment & Evaluation in Higher Education, 36*(5), 509-547. doi: 10.1080/02602930903540991.

Trenary, A., & Farrar, H. (2016). Use of a professional writing rubric as a teaching strategy to improve scholarly writing. *The Oklahoma Nurse,* 12-13.

Tschannen-Moran, M., & Johnson, D. (2011). Exploring literacy teachers' self-efficacy beliefs: Potential sources at play. *Teaching and Teacher Education*, *27*(4), 751-761.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263-287.

Weigle, S. C. (2007). Teaching writing teachers about assessment. *Journal of Second Language Writing, 16*(3), 194-209. doi:10.1016/j.jslw.2007.07.004

Wolfe, E. W., Mathews, S., & Vickers, D. (2010). The effectiveness and efficiency of distributed online, regional online, and regional face-to-face training for writing assessment raters. *The Journal of Technology, Learning and Assessment, 10*(1).

Zimmerman, B. S., Morgan, D. N., & Kidder-Brown, M. K. (2014). The use of conceptual and pedagogical tools as mediators of preservice teachers' perceptions of self as writers and future teachers of writing. *Action in Teacher Education, 36*(2), 141-156.