

Development and Initial Validation of the Early Elementary Writing Rubric to Inform Instruction for Kindergarten and First-Grade Students

Assessment for Effective Intervention
2022, Vol. 47(4) 220–233
© Hammill Institute on Disabilities 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/15345084211065977
aei.sagepub.com



Meaghan McKenna, PhD¹ , Robert F. Dedrick, PhD¹,
and Howard Goldstein, PhD¹ 

Abstract

This article describes the development of the Early Elementary Writing Rubric (EEWR), an analytic assessment designed to measure kindergarten and first-grade writing and inform educators' instruction. Crocker and Algina's (1986) approach to instrument development and validation was used as a guide to create and refine the writing measure. Study 1 describes the development of the 10-item measure (response scale ranges from 0 = *Beginning of Kindergarten* to 5 = *End of First Grade*). Educators participated in focus groups, expert panel review, cognitive interviews, and pretesting as part of the instrument development process. Study 2 evaluates measurement quality in terms of score reliability and validity. Data from writing samples produced by 634 students in kindergarten and first-grade classrooms were collected during pilot testing. An exploratory factor analysis was conducted to evaluate the psychometric properties of the EEWR. A one-factor model fit the data for all writing genres and all scoring elements were retained with loadings ranging from 0.49 to 0.92. Internal consistency reliability was high and ranged from .89 to .91. Interrater reliability between the researcher and participants varied from poor to good and means ranged from 52% to 72%. First-grade students received higher scores than kindergartners on all 10 scoring elements. The EEWR holds promise as an acceptable, useful, and psychometrically sound measure of early writing. Further iterative development is needed to fully investigate its ability to accurately identify the present level of student performance and to determine sensitivity to developmental and instruction gains.

Keywords

writing, assessment, kindergarten, first grade

Although the ability to express one's thoughts in writing is critical for success in school and the workplace, only 25% of students are proficient writers by the 8th to 12th grades (National Center for Education Statistics, 2012). Despite the development of tools to identify young students at risk of reading difficulties (e.g., Kaminski & Good, 1998), there are only a few measures with emerging evidence of validity to assess kindergarten and first-grade writing (Coker & Ritchey, 2014; Keller-Margulis et al., 2019). Writing assessment research rarely occurs in primary grades (Philippakos & FitzPatrick, 2018), especially in kindergarten. Because teachers find it difficult to assess writing (Feenstra, 2014), writing difficulty often goes undiscovered (White, 2013). The classroom teacher has a critical role in identifying students who have difficulty responding to core instruction (Bondie et al., 2019). Accurate measures of writing development are needed to plan for differentiated core instruction and tiered interventions to reduce academic difficulties in children at risk (Coker

& Ritchey, 2014). Educators need a valid and reliable measure of writing to assess student response to instruction (Lembke et al., 2010); to identify which skills are mastered, emerging, or absent; and to maximize the effectiveness of instruction (McMaster et al., 2017).

Technical Adequacy of Assessments

All assessments must meet technical requirements of validity and reliability (American Educational Research Association [AERA] et al., 2014; Romig et al., 2017).

¹University of South Florida, Tampa, USA

Corresponding Author:

Meaghan McKenna, Department of Educational Psychology, University of Connecticut, 249 Glenbrook Road, Unit 3064, Storrs, CT 06269, USA.
Email: meaghan.mckenna@uconn.edu

Associate Editor: Nathan Stevenson

Content validity refers to how accurately items on an assessment represent the construct (e.g., writing). *Response process validity* is the analysis of the responses of individual respondents or observers. *Internal structure validity* is the degree to which items fit the underlying constructs. *Relations to other variables* include positive or negative relations between scores related to the construct. *Consequences* refer to the evaluation of the intended and unintended effects of an assessment. *Criterion validity* assesses how a measure correlates with another measure (Taylor, 2006) but is often difficult for writing due to the differences in the field of how much weight should be placed on various elements. While reliability is necessary, it is not sufficient because a measure that produces accurate but invalid scores is meaningless (Romig et al., 2017). Reliability provides an estimate of the upper limit on validity, and it is a more sensitive index of measurement quality (Conway et al., 1995). *Internal consistency reliability* is calculated to check on the quality of the data and how well it is measuring the intended construct (McCrae et al., 2011) and *interrater reliability* is the extent of agreement among scorers.

Types of Writing Assessments

Although there are different types of writing assessments available, many have serious shortcomings. Holistic and primary trait (Lloyd-Jones, 1977) scoring involves assigning a rating on a rubric based on the overall quality of the writing piece. Performance on several criteria (e.g., organization and spelling) is associated with one score that makes these rubrics quick to use (Huot, 2002). However, one score leads to broad interpretation and does not readily inform instruction because specific strengths and weaknesses are not identified (Coker & Ritchey, 2010; Lam, 2018). Concerns with interrater reliability and criterion validity for holistic rubrics exist (Espin et al., 2004). When using primary trait scoring, raters are provided guides customized to the topic or prompt associated with the writing task (Applebee, 2000) along with performance aspects considered important for success and examples of writing that illustrate successful task completion (Frey, 2018). Primary trait scoring can be implemented with reliability (Espin et al., 2004) but development of new scoring materials for each task is time intensive (Frey, 2018).

Analytic assessment involves analyzing writing using a set of elements to determine proficiency (Coker & Ritchey, 2010) and allows for comprehensive measurement (Espin et al., 2004). The *6+1 trait scoring system* offers an example. Seven elements are scored: idea development, organization, word choice, sentence fluency, voice, presentation, and conventions (Northwest Regional Educational Laboratory [NREL], 2014). Educators refer to *voice* as the students' use of individual expression. However, definitions of voice

vary, and it is difficult to measure quantitatively and score reliably (Jeffery, 2011). The 6+1 trait rubric is seven pages and consequently time-consuming to use. Empirical evidence is limited (Gansle et al., 2006). Kim et al. (2014) identified the dimensions of first-grade narrative writing using an adapted 6+1 trait scoring system, a two-factor model yielded a good fit with four ideation (main ideas, organization, word choice, sentence fluency) and three transcription (spelling, handwriting, and mechanics) items.

Writing curriculum-based measurements (CBMs) are multiple probes of equivalent difficulty that can be administered repeatedly, yielding time series data that reflect student progress across a variety of metrics (e.g., correct writing sequences). Reliability ranges from .70 to .90 (moderately to relatively strong; McMaster & Espin, 2007). While CBMs are quick to administer and are designed to indicate global performance, the general domain nature is insufficient to guide instruction (Puranik et al., 2020). McMaster and Espin (2007) reported that it is difficult to determine if the simple indices measured (e.g., total words written) are sufficient for identifying elementary students at risk of meeting grade-level standards. The complicated interpretation procedures have led teachers to report lower usability and a need for ongoing scoring training as compared with reading CBMs (Payan et al., 2019). To date, Payan conducted the only study where teachers scoring writing CBMs shared perceptions. Allen et al. (2020) noted a need for teachers to score writing CBMs to examine feasibility.

Purpose

A wider variety of validated assessments of writing beyond the more common CBMs are needed, but reliability and validity studies have been sparse (Gansle et al., 2006). In light of the limitations of existing assessments of writing, there is a need for a psychometrically sound instrument that contains clear criteria and covers the breadth of development displayed by kindergarten through first-grade students. We developed the Early Elementary Writing Rubric (EEWR) for measuring kindergarten and first-grade writing. This rubric was developed to meet measurement (e.g., AERA et al., 2014) and educational standards. This rubric is (a) grounded in the Simple View of Writing Theory (Juel et al., 1986) and (b) able to measure all writing genres instructed, as standards for kindergarten and first-grade children include knowledge of text types and purposes (i.e., opinion, informational, and narrative writing). The Simple View of Writing (Juel et al., 1986) involves the component skills of ideation (organization, topic maintenance, sentence structure, and vocabulary) and transcription (handwriting, spelling, punctuation, and upper/lower case letter use). Ideation is related to oral language and

includes the planning, generation, and organization of writing. Transcription denotes handwriting and spelling (Kim et al., 2015). The rubric was designed to be fair, authentic, align with input from practitioners, and provide teachers with information about students' progress in meeting grade-level expectations (National Commission on Writing, 2006).

The aims of this study were twofold. In Study 1, we developed the EEWR, an assessment designed to evaluate beginning writing accurately and efficiently. We followed the Crocker and Algina (1986) approach to instrument development and validation. Content validity was examined to ensure adequate coverage of items in relation to the writing construct (Messick, 1975). In Study 2, relations to other variables and internal structure validity were investigated to further refine and support the psychometric quality of this assessment. Relations to other variables involved examining age and writing genres, as differences in the writing produced by kindergarteners and first graders were compared as well as three different genres (McCoach et al., 2013). Previous studies have found differences between narrative and expository writing at both the microstructure and the macrostructure for students in later elementary grade levels (Hall-Mills & Apel, 2015). The effect of genre on writing proficiency requires more research (Jeong, 2017), especially in early elementary grades. Internal structure validity informs the use of an assessment and determines how items are grouped together (Rios & Wells, 2014). Internal consistency, interrater reliability, and acceptability of this assessment were evaluated.

Study 1: Construction of An Instrument Assessing Early Elementary Writing

The EEWR was developed through an iterative, multistep process to generate. Crocker and Algina (1986) recommend systematic steps for developing measures. Content validity was assessed by gathering information aligned with the following steps: (a) identification of the purpose of the instrument (Step 1); (b) creation of a blueprint and generation of an initial item pool (Steps 2–4); (c) expert panel review (Step 5); (d) cognitive interviews (Step 5); and (e) pretesting (Step 5). Pilot testing (Step 6) was conducted in Study 2.

Identification of the Purpose of the Instrument

As noted in the introduction, a measure to identify specific strengths and weaknesses to assist with planning core instruction and tiered interventions is needed. The first author worked with elementary teachers to create the initial version of the rubric used during an intervention study (McKenna et al., 2021). Teachers identified the most important skills for assessing writing in kindergarten and

first grade that matched the component skills in the Simple View of Writing.

Creation of Blueprint and Item Pool

Five focus groups were held to assist with creating a test blueprint and item pool. A combination of kindergarten and first-grade teachers, resource teachers and coaches, and an assistant principal participated in each focus group with attendance ranging from 5 to 12 participants. Focus groups lasted approximately 1 hr and were audio-recorded. An outline of activities containing points of discussion was used to ensure consistency in the structure of each group. Participants provided feedback on (a) instrument appearance, (b) skills assessed, (c) item criteria and clarity, (d) response categories, and (e) scoring instructions. First, participants reviewed the eight writing elements: handwriting, punctuation, uppercase/lowercase letters, grammar, punctuation, vocabulary, on topic, and organization on the first version of the rubric found in online supplemental Table S1. They suggested skills be removed or added, how skills could be quantified at the beginning, middle, end of kindergarten and first grade, and how descriptions could be written. Participants each had access to the Florida Department of Education *English Language Arts Standards Implementation Guide*, a document designed to guide vertical planning within and across grade levels. The standards were used to assist with identifying the skills assessed and develop response categories. The stages of Krueger's (1994) systematic framework analysis were followed. The recordings of the focus groups were transcribed during the familiarization stage. During the next stage, the relevant pieces of the transcript were sorted under the writing element they represented to create a thematic framework. Responses to each of the questions (e.g., appearance of the rubric and scale) were categorized to assist with indexing and sorting the data so similar responses were grouped together and revisions could be made to the rubric in the final stage. The resulting revised rubric was submitted to the expert panel for review (see online supplemental Table S2).

Expert Panel

A total of 13 experts who authored publications in writing, worked for a department of education, or held supervisory roles in a school district's teaching and learning department provided feedback on the rubric, and 11 completed a questionnaire in Qualtrics. None of the experts participated in the focus groups. Experts were emailed the rubric and asked questions addressing (a) adequacy and clarity of content coverage, (b) relevance of the item content for the proposed instrument, (c) wording and structure of the items, and (d) appropriateness of the response scale. A 5-point scale was used to indicate the relevance of each element when

measuring kindergarten and first-grade writing. The expert panel's relevance ratings were examined using the item content validity index (I-CVI; Lynn, 1986). The I-CVI was calculated by taking the number of experts who provided ratings of relevant or very relevant for an item divided by the total number of experts who rated an item. I-CVI indexes ranged from 82% to 100% for all nine elements rated: handwriting, spacing, punctuation, uppercase/lowercase letters, grammar, punctuation, vocabulary, on topic, and organization. Expert input on criteria and feasibility for scoring was used to make further revisions and prepare for cognitive interviews.

Cognitive Interviews

Fifteen kindergarten and first-grade teachers, resource teachers, and coaches who did not participate in focus groups completed cognitive interviews (Willis, 2004). Each participant was given a copy of a writing sample and a scoring rubric (see Table S3). After the participant read the writing sample, a think aloud occurred as they used the rubric to score each writing element. This provided information on how participants: (a) worked through the items, (b) assessed if a response category was appropriate for the item, and (c) selected each score. Items on the rubric that appeared confusing were discussed along with how they could be revised. For example, additions were made to the vocabulary scale to allow for the progression to represent each tier of vocabulary: basic concepts and words heard often (Tier 1); high frequency, high utility words requiring direct instruction (Tier 2); and concepts limited to specific content/domains (Tier 3; Beck et al., 2013). Participants also suggested that the sentence structure scoring element be divided into two scales: one to measure grammar and the other to measure variety. An educator also suggested including the standards corresponding to each of the scoring elements. Further revisions to the rubric were made following the cognitive interviews to prepare for pretesting.

Pretesting

Forty-nine teachers pretested the rubric at their school sites with attendance at each meeting ranging from 2 to 10 participants. Pretesting allowed for assessment of reliability and discussion of any items that remained unclear before a full-scale study. Participants were provided with the scoring manual and previously collected writing samples. A 20-min overview on how to use the rubric and a corresponding manual to assist with scoring each element was provided. A cover sheet specific to scoring each element was in the manual as well as pictures of writing samples with a score for each element and an explanation of why that score was received. The cover sheet for sentence

structure reminded teachers to count the number of sentences attempted when scoring for grammar and variety. This helped them to determine if the sentence was complete and made sense, and if they would say the sentence the same way when scoring grammar and to determine how many sentences began differently and contained a different number of words when scoring variety. For example, the explanation for a score of 5 for vocabulary says: uses several words that require explicit instruction and are related to content (savanna, male, female, mane, travel, north, blend, leap, and attack).

All teachers used a rubric to score the same writing sample. Then scores were compared. Disagreements in scoring were discussed and the criteria were reviewed for elements along with what may have needed to be present or absent from the writing piece so that all raters were aware of what to look for when scoring each of the subsequent writing pieces. For example, if a teacher scored a 5 for punctuation because they noted punctuation for every sentence, but only periods were used, we discussed that the highest score that could be obtained was a 4 because there was no variety (i.e., periods were used throughout the entire writing piece) and there were commas missing from a list of items. Two of the groups had an interrater agreement that fell below 80%. As a result of the high percentage of disagreement, an additional sample was scored. The scores given to the second sample were discussed and agreement was at or above 80% when calculating the agreement of the scores given to each scoring element (e.g., letter formation and grammar) on the rubric. Several teachers recommended having a one-page cover sheet added to the scoring manual with tips for scoring the writing samples. They also shared how procedures for collecting writing samples could be revised before beginning Study 2.

Study 2: Validation, Reliability, and Acceptability Study

After the EEWR was developed, the next study was conducted to validate the rubric, assess reliability, and determine educators' perceptions of the acceptability of the rubric. The following research questions were examined:

Research Question 1: Are there differences in the scores received by kindergarten and first-grade students?

Research Question 2: What is the internal structure validity of the scores from the rubric?

Research Question 3: Does the factor structure differ across writing genres?

Research Question 4: What is the internal consistency reliability of the scores from the resultant factors?

Research Question 5: What is the interrater reliability of the scores from the rubric?

Table 1. Student Demographic Information and Assessment Performance.

Demographic Characteristics	Total (<i>n</i>)	Free	Reduced	Non-free or reduced
Kindergarten	285	108 (37.9%)	10 (3.5%)	167 (58.6%)
First	349	126 (36.1%)	13 (3.7%)	210 (60.2%)
Individualized Education Program	80	31 (38.8%)	6 (7.5%)	43 (53.8%)
504	4	3 (75%)	0 (0%)	1 (25%)
English language learner	76	57 (75%)	6 (7.9%)	13 (17.1%)
Retention	10	7 (70%)	0 (0%)	3 (30%)
White	326	43 (13.2%)	5 (1.5%)	278 (85.3%)
Black	91	78 (85.7%)	4 (4.4%)	9 (9.9%)
Hispanic	170	98 (57.6%)	12 (7.1%)	60 (35.3%)
Asian	14	1 (7.1%)	0 (0%)	13 (92.9%)
Multiracial	33	14 (42.4%)	2 (6.1%)	17 (51.5%)
Female	293	108 (36.9%)	6 (2%)	179 (61.1%)
Male	341	126 (37%)	17 (5%)	198 (58%)
iReady fall percentile on level	524	153 (29.2%)	18(3.4%)	353 (67.4%)
iReady fall percentile below level	109	80 (73.4%)	5 (4.6%)	24 (22%)
iReady winter percentile on level	509	145 (28.5%)	14 (2.8%)	350 (68.7%)
iReady winter percentile below level	123	88 (71.5%)	8 (6.5%)	27 (22%)

Research Question 6: Do teachers find this rubric to be acceptable for measuring student writing?

Method

Participants

A total of 634 students (285 kindergarteners and 349 first graders), 55 classroom teachers (25 kindergarten and 30 first grade), 1 site-based administrator, and 5 educators serving as a coach or resource teacher participated in Study 2. Diagram S4 in the online supplemental materials indicates the number of participants throughout the study, as 13 teachers dropped out because of other obligations, illness, or trouble meeting deadlines. All educators were female. The majority held a bachelor's degree. Their years of experience working in a school-based setting ranged from 1 to 42. Table S5 in the online supplemental materials contains detailed educator demographic information. A broad representation of students from Title I and non-Title I schools, geographically dispersed across a large district in Florida, participated. Table 1 contains detailed student demographic information.

Pilot Testing

The initial pilot of the instrument provided preliminary evidence of the factor structure to be established (McCoach et al., 2013). A one-page handout was included in each participant's binder to remind them of procedures. The handout contained 4-steps to follow: (a) getting the envelope with the writing assessment materials, (b) playing a video, (c)

setting a stopwatch to record how long students spent writing, and (d) collecting the writing samples. Prior to collecting the writing samples, the teachers received three envelopes labeled with the genre (e.g., opinion), date, and time of the assessment. The envelopes contained the materials needed for collecting the writing samples. Information on how to access the videos was emailed to the teachers. On the day of each writing assessment, the teacher played the video corresponding to the writing genre to be collected. The video provided all directions for the students. The videos gave teachers cues of when to pause so that papers could be distributed to students after they selected the prompt that they wanted to write about. Students were given 2 min to plan for their writing on the blank paper. Then they were given up to 15 min to write on their lined paper. The teachers set a stopwatch to track how much time each student spent independently writing on their lined paper. The video provided a reminder for when the time was almost up. It also alerted students and teachers when writing stopped. Teachers then collected and labeled the papers.

Teachers followed the calendar of testing dates, times, and the genre to administer to collect the writing samples in 2 weeks in January. The order of the three genres was counterbalanced across 12 school sites (6 Title I and 6 Non-Title I). The order at each of the sites was selected randomly. The kindergarteners and first graders had the opportunity to select one of two prompts presented for each genre. The prompts were identified based on feedback from the focus groups and cognitive interviews. The prompts identified for narrative were "Write a story about a friend" and "Write a story about a time that you were scared"; for informational were "Write about an animal" and "Write

about how to brush your teeth”; and for opinion were “Do you like playing inside or outside?” and “Would you rather go to a pool or the beach?”

There was at least one fidelity check per classroom for a total of 97 (59%) of 165 sessions. The researcher, coaches, resource teachers, and speech-language pathologists (SLPs) working at schools where the study took place filled out a checklist found in online supplemental Table S6 to ensure the assessment was administered accurately. Assessments were administered with 100% fidelity.

After the assessment was administered, teachers scored samples for 25% to 30% of the students randomly selected from their class. The rubric used to score writing can be found in online supplemental Table S7. If a student was absent, they scored the sample of an alternate student. A one-page hand-out contained seven steps for teachers to follow: (a) getting the envelope and binder, (b) recording scores on a rubric and using the manual to assist as needed, (c) placing labels on the completed rubrics and storing them in the binder, (d) exchanging writing samples with a teacher partner, (e) scoring writing samples from a partner’s classroom, (f) placing labels on completed rubrics and storing them in the binder, and (g) returning writing samples to a partner teacher and collecting back writing samples from the students in their classroom. The classroom teacher was aware of the students in their classroom because a list of the student names and unique identifiers were provided to assist with labeling papers and removing names. Each teacher scored 25% to 30% of the writing samples of students from another classroom. The teachers were unaware of who the students were when scoring the writing samples from the other classroom because names had been removed. These data were used for interrater reliability analyses.

Pilot Testing Results

Descriptive statistics for each of the elements for all writing samples and the writing samples separated by grade level are presented in online supplemental Table S9. The researcher scored a total of 1,733 writing samples (563 narrative, 581 informational, and 589 opinion). There was a total of 786 writing samples (253 Narrative, 265 Informational, and 268 Opinion) produced by 285 kindergarteners and 947 writing samples (310 Narrative, 316 Informational, and 321 Opinion) produced by 349 first graders. Students produced up to 3 writing samples: one per genre. Means were computed by using a 6-point scale (*Beginning of Kindergarten—End of First Grade*) ranging from 0 to 5 for each scoring element organized according to the level of proficiency.

All Writing Samples

Students’ narrative, informational, and opinion samples received the full range of scores from 0 to 5 for all elements

with the exception of vocabulary for the narrative and opinion genres for which scores ranged from 0 to 4. The means for the scoring elements ranged from 1.36 to 3.31 for narrative, 1.40 to 3.31 for informational, and 1.23 to 3.41 for opinion writing. Spacing was always the element with the highest mean. Variety had the lowest mean for narrative and opinion writing. The organization had the lowest mean for informational writing. When comparing the means obtained for each element across genres, they were all relatively close, with differences ranging from 0.04 to 0.46. The element with the lowest difference across genres was letter formation. The element with the highest difference across genres was on topic; the mean for narrative writing scores was higher than the mean obtained for opinion writing scores. Skewness ranged from (−0.74 to 1.35) and kurtosis ranged from (−1.70 to 1.40). These values suggested that there were no severe departures from normality (Curran et al., 1996). The standard deviation ranged from (0.86–1.80) and indicated low to high dispersion among the scores.

Writing Samples Separated by Grade Level

The full range of scores 0 to 5 was reflected in all kindergarten writing samples for letter formation, spacing, upper/lowercase letters, and punctuation for all three genres. The scores for kindergarteners did not always reach the maximum. For example, high scores were 3 for organization and 4 for spelling. Grammar reached a maximum of 4 for narrative and informational but 5 for opinion. Variety reached a maximum of 5 for narrative and informational and 4 for opinion. Vocabulary ranged from 0 to 3 for narrative, 0 to 5 for informational, and 0 to 4 for opinion. On Topic ranged from 0 to 4 for informational and 0 to 5 for narrative and opinion. The full range of scores (0–5) was demonstrated in first-grade writing for all elements except vocabulary for the narrative and opinion genres (i.e., 0–4) and letter formation for narrative (i.e., 1–5).

The means of the scores for kindergarteners ranged from 0.71 to 2.72 for narratives, 0.82 to 2.80 for informational writing, and 0.76 to 2.91 for opinion writing. Spacing was the element with the highest mean and variety was the element with the lowest mean for all genres. The means of the scores for first graders ranged from 1.88 to 3.80 for narratives, 1.63 to 3.82 for opinion writing, and 1.76 to 3.75 for informational writing. Spacing was the element with the highest means for all genres. Grammar was the element with the lowest mean for narratives, the organization was the element with the lowest mean for informational writing, and variety was the element with the lowest mean for opinion writing. Skewness ranged from −0.36 to 1.26 for kindergarteners and 0.90 to −1.17 for first graders. Kurtosis ranged from −1.48 to 3.27 for kindergarteners and −1.45 to 1.33 for first graders. These values suggested that there were no severe departures from normality (Curran et al.,

1996), with the exception of the kurtosis value of 3.27 for spelling for kindergarten informational writing. The standard deviation ranged from 0.61 to 1.77 for kindergarteners and 0.67 to 1.66 for first graders, which indicated low to high dispersion.

Online supplemental Figure S8 presents the means and standard deviations of each scoring element when using data gathered from all writing samples. When the rubric was developed, the six scores were meant to reflect teacher expectations for the beginning, middle, and end of kindergarten through first grade. Thus, kindergarteners were anticipated to receive a score of at least 1, and first-grade students were anticipated to receive a score of at least 4. Means of kindergarten student scores met or exceeded that expectation for 8 of the 10 elements. However, the means obtained by first graders fell below a priori expectations on all elements. The first graders had higher means on all elements compared with the kindergarteners. The only element with significant overlap between the grades was letter formation. Independent samples *t* tests were conducted to determine differences between the mean scores obtained by kindergarteners and first graders. Statistically significant differences ($p = .000$) resulted for all elements when the data from all writing genres combined, narrative, informational, and opinion writing were analyzed. Cohen's *d* (Cohen, 1988) was computed using the differences in the means divided by the pooled standard deviation. Cohen's *d* ranged from 0.55 to 1.06, with an average of 0.82. The elements of Grammar, Variety, Spelling, Vocabulary, On Topic, and Organization had large effects. The elements of Letter Formation, Spacing, Upper/lowercase Letters, and Punctuation had medium effects. Frequencies for all scores can be found in online supplemental Table S10.

Internal Structure Validity

A total of four exploratory factor analyses (EFAs) were conducted for: (a) all genres, (b) narrative writing, (c) informational writing, and (d) opinion writing. EFAs were conducted using Mplus 8.3 (Muthén & Muthén, 2019) using maximum likelihood estimation and oblique rotation. When determining the number of factors, a review of eigenvalues, scree plots, and parallel analyses were used to indicate the number of factors to retain. Eigenvalues were obtained and there were two values above 1.0 for all genres, 5.50 and 1.09; two values above 1.0 for narrative writing, 5.80 and 1.06; two values above 1.0 for informational writing, 5.41 and 1.08; and two values above 1.0 for opinion writing, 5.50 and 1.14. There was a large difference between the two eigenvalues, one much higher than 1.0 and the other close in value to 1.0. A review of scree plots and parallel analyses indicated that there was only one factor to retain.

A cutoff of 0.35 was used to determine if an item loaded on the one factor (Henson & Roberts, 2006). All loadings

exceeded this criterion. One-factor solutions resulted for each EFA. This factor represents the construct of early elementary writing. When reviewing the two-factor models, elements potentially loading on a second factor were grammar and variety for informational, narrative, and all genres. There was no convergence for opinion writing. The pattern coefficients (loadings) of items can be found in online supplemental Table S11. The one-factor solution produced similar results for all EFAs, accounting for a total variance of 55% for all genres; 58% for narrative; 54% for informational; and 55% for opinion. Loadings ranged from 0.53 to 0.87 for all genres, 0.53 to 0.92 for narrative, 0.49 to 0.85 for informational, and 0.52 to 0.89 for opinion. Internal consistency reliability estimates (as measured by Cronbach's α) are as follows: all genres: $\alpha = .89$; narrative: $\alpha = .91$; informational $\alpha = .89$ and opinion $\alpha = .89$.

Interrater Reliability

Interrater reliability was conducted for at least 25% of the 9 to 26 writing samples scored by all 55 participants. Reliability was calculated by dividing the number of agreements by the total number of ratings. Fleiss' κ coefficient was calculated using SPSS Statistics version 26. Table 2 contains means and standard deviations of interrater agreement percentages, one-point discrepancy agreement percentages, Fleiss' κ , and Krippendorff's α reliability estimates collected overall and for all skills measured. Online supplemental Table S12 contains each participant's data.

Exact match scoring agreement percentages between each participant and the researcher for all ratings ranged from 48% to 80%. The overall exact match agreement average was 65%. Exact match agreement for all 10 of the scoring elements ranged from 20% to 100% for Letter Formation, 22% to 92% for Spacing, 17% to 100% for Upper/Lowercase Letters, 33% to 91% for Grammar, 29% to 94% for Variety, 17% to 96% for Punctuation, 24% to 100% for Spelling, 31% to 95% for Vocabulary, 18% to 94% for On Topic, and 29% to 94% for Organization. The organization had the highest exact match agreement average of 72% and Spelling had the lowest, 52%.

The 1-point discrepancy scoring agreement between raters was calculated and ranged from 83% to 99% for all ratings. The overall 1-point discrepancy scoring agreement average was 91%. The 1-point discrepancy ranged from 67% to 100% for Letter Formation, 56% to 100% for Spacing, 56% to 100% for Upper/Lowercase Letters, 59% to 100% for Grammar, 63% to 100% for Variety, 50% to 100% for Punctuation, 53% to 100% for Spelling, 80% to 100% for Vocabulary, 67% to 100% for On Topic, and 71% to 100% for Organization. Vocabulary had the highest 1-point discrepancy scoring agreement average of 97% and spelling had the lowest, 81%.

Table 2. Interrater Reliability Data.

n	Overall		Letter formation		Spacing		Upper/Lowercase letters		Grammar		Variety		Punctuation		Spelling		Vocabulary		On topic		Organization	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
56	.65%	.08	.66%	.15	.64%	.14	.59%	.17	.69%	.13	.63%	.15	.71%	.18	.52%	.16	.66%	.16	.64%	.16	.72%	.14
56	.91%	.04	.96%	.08	.93%	.07	.83%	.12	.93%	.09	.89%	.11	.90%	.10	.81%	.13	.97%	.04	.95%	.07	.96%	.06
56	.54	.09	.52	.19	.50	.19	.46	.20	.52	.19	.43	.20	.57	.23	.35	.20	.46	.23	.48	.20	.56	.21
43	.77	.10	.67	.28	.67	.31	.55	.27	.60	.28	.54	.29	.74	.23	.50	.26	.59	.28	.65	.23	.62	.27

Exact interrater agreement averages

I-point discrepancy averages

K averages

Krippendorff's α

Fleiss' κ ranged from .34 to .74 indicating a range of fair to substantial agreement on all items scored (Landis & Koch, 1977). The Fleiss' κ average overall was .54 meaning moderate agreement. Fleiss' κ for all 10 of the scoring elements ranged from: $-.02$ to 1.0 for Letter Formation, $.06$ to 0.87 for Spacing, 0.06 to 1.0 for Upper/Lowercase Letters, $.12$ to 0.89 for Grammar, $.01$ to 0.88 for Variety, $-.03$ to 0.94 for Punctuation, $.01$ to 1.0 for Spelling, $-.23$ to 0.91 for Vocabulary, $-.02$ to 0.91 for On Topic, and $-.04$ to 0.91 for Organization. A range of poor to almost perfect agreement was presented for all 10 scoring elements. The Fleiss' κ averages were moderate with the exception of Spelling which would be interpreted as fair.

Because three different raters scored writing samples, the KALPHA macro (Hayes & Krippendorff, 2007) was used to compute Krippendorff's α to obtain reliability estimates for each rating trio. The KALPHA macro was used to compute Krippendorff's α because judgments can be made with any number of observers, with or without missing data. A total of 43 rating trios scored between 6 and 24 of the same writing samples. Krippendorff's α ranged from $.48$ to $.91$ for all ratings; 20 out of the 43 trios (47%) obtained an α at or above $.80$. The Krippendorff's α average for all scores was $.77$. Krippendorff's α ranged from: $-.02$ to 1.0 for Letter Formation, $-.28$ to 1.0 for Spacing, $-.03$ to $.98$ for Upper/Lowercase Letters, $-.03$ to 1.0 for Grammar, $-.11$ to 1.0 for Variety, $-.08$ to 1.0 for Punctuation, $-.20$ to $.87$ for Spelling, $-.18$ to 1.0 for Vocabulary, $.04$ to $.96$ for On Topic, and $-.14$ to $.97$ for Organization. Spelling had the lowest Krippendorff's α average of $.50$.

Social Validity Assessment

Forty-seven participants completed a social validity survey (online supplemental Table S13). They provided ratings ranging from strongly agree to strongly disagree for 18 questions about administrative procedures and usability of the rubric and manual. Three of the items, overall ease of scoring, scoring spelling, and scoring vocabulary, received a neutral rating of at or above 2.50, whereas ratings for all other items fell between strongly agree and agree. Participants also responded to three short answer questions. Thirty-five percent of teachers indicated that the scoring manual was helpful because of the examples and guidelines. Fifty-six percent of teachers said the rubric allowed for assessment of a variety of aspects of writing, and the breakdown of each scoring element was useful for planning because decisions could be made on skills to focus on during whole group, small group, and individual instruction. Twenty-one percent of teachers noted that the objectivity of the rubric allowed for consistency in report card grading. Thirty-one percent of the teachers said that it was time-consuming; however, 37% of these educators also noted that once they had more opportunities for scoring, the rubric

would become easier to use and one of these teachers acknowledged that all the scoring elements were important for kindergarten and first-grade writing and should remain on the rubric. Forty-four percent of the teachers said the progressions for scoring elements were well organized, clear, explicit and that the rubric was easy to use. Forty-seven percent of teachers noted that a few scoring elements required further revision, with 88% reporting that spelling and vocabulary were the most challenging to score.

The teachers appreciated having opportunities to meet with this researcher throughout the instrument development process. They felt that assessment procedures were well organized, but modifications could be made for future administrations. They suggested using a different method for student prompt selection, more time for planning, additional information in the videos to provide students with the meaning of each genre and a reminder for how to structure their writing (e.g., use supporting details to tell me why you think or feel that way), and a checklist for students to remind them of what they were being assessed on when producing their writing.

Discussion

This article describes the development and initial validation of the EEWR. Consistent with the Standards for Educational and Psychological Testing (AERA et al., 2014), multiple sources of validity and reliability evidence were collected to support the intended use of the rubric. There were five major findings: (1) The rubric distinguished performance and developmental differences of kindergarteners versus first graders providing evidence of relations to other variables. (2) The results of EFAs indicate that kindergarten and first-grade narrative, informational, and opinion writing is characterized by a one-factor model. (3) High internal consistency estimates provide evidence for the internal consistency of this latent construct. (4) The interrater reliability agreement analyses revealed modest agreement in teachers' ratings of student writing, a result that will require continued refinements to the measure and more extensive training of teachers. (5) The social validity feedback provided was useful for guiding revisions and planning for future research to further refine the rubric.

As predicted, first graders had higher means on all element scores when compared with kindergarteners. This provides evidence for the validity of the rubric because it was sensitive to expected developmental differences between grade levels. When reviewing the scores, none of the participants received the highest scores on all elements assessed. Thus, the rubric was relatively free of floor or ceiling effects. The rubric was developed with a progression that reflected expectations of beginning of kindergarten to end of first grade. The frequency data obtained indicates that the majority of kindergarteners exceeded

expectations in letter formation and spacing; fell within the acceptable range for upper/lowercase letters and spelling; fell slightly below expectations for vocabulary, on topic, and organization; and fell below expectations for grammar, variety, and punctuation. The majority of the first graders fell within acceptable ranges for letter formation and spacing but below expectations for upper/lowercase letters, grammar, variety, punctuation, spelling, vocabulary, on topic, and organization. These results indicate promise for using this rubric to also score the writing of second-grade students.

The majority of the teacher participants commented that this rubric helped them plan for instruction and address parts of writing that had not previously been taught. According to the National Assessment of Educational Progress, only 28% of fourth-grade students write proficiently (National Center for Education Statistics, 2012). International assessment results (e.g., United Kingdom Department for Education, 2012) indicate that students worldwide have poor writing skills. Setting learning goals using criteria within this measure and providing instruction to meet these goals has the potential to produce well-equipped young writers. High expectations in early elementary grades may be needed to set up learners for success in writing. These expectations appear reasonable because there were some kindergarteners and first graders who successfully met or exceeded expectations on this measure.

Internal Structure Validity

The internal structure is likely to stand up to further scrutiny, given the relatively large sample size. With such a large sample, one might expect to find relatively stable loadings when replicated in future studies (MacCallum et al., 1999). There were no differences in the EFAs when comparing the results from all the genres combined, narrative writing, informational writing, and opinion writing. The loadings for every EFA were well above the established cut score, and the loadings for each of the scoring elements were relatively close in value.

The nature of writing continues to be open to debate. Gansle et al. (2006) assessed story writing in first through fifth grade and suggested that six trait ratings contributed to a single dimension. Kim et al. (2014) used an adapted 6+1 trait scoring system to assess the narrative writing of first graders, but their results revealed a two-factor model. In contrast, Puranik et al. (2020) found that a system similar to Kim et al. was best modeled as unidimensional for kindergarteners. Wagner et al. (2011) found writing of first and fourth graders to be characterized by a 5-factor model: macro-organization, complexity, productivity, spelling and punctuation, and handwriting fluency using the data from a variety of quantitative measures (e.g., number of different words and mean length of the T-unit). It is worth noting that

these models were derived using scores obtained on a variety of assessments (e.g., Wechsler Individual Achievement Test—Third Edition). Past studies often have explored one (e.g., Kim et al., 2011) or two dimensions (e.g., Olinghouse, 2008) of writing productivity and quality. However, these studies used norm-referenced (e.g., Test of Language Development—Intermediate) and researcher-generated measures (e.g., handwriting automaticity, advanced planning scale).

The factor loadings and unidimensional construct of writing that resulted from this study differed from the Simple View of Writing Theory used to guide assessment development. These findings support that the skills represented by elements on the EEWWR can be addressed separately or grouped together. The correlations from each of the skills ranged from .33 to .94. Grammar and variety (.87–.94) were highly correlated for narrative, opinion, informational, and all genres. Grammar and variety are both part of the sentence structure scoring element. On topic and organization (.82–.86) were highly correlated for narrative, opinion, and all genres. Vocabulary and on topic (.83) were highly correlated for the narrative genre. The only skills that were highly correlated for every genre assessed were grammar and variety. Practitioners involved in this study found that having these as separate scoring elements was useful when measuring student writing. The practitioner providing instruction can determine the most appropriate skill(s) to address during whole group, small group, or individual instruction based on data. In an analogous study, Lonigan and Milburn (2017) explored the dimensions of the language construct and determined there are fewer dimensions than often proposed or found in the various language assessment subtests. The two dimensions that resulted in their study were syntax and vocabulary. Although a one-factor model resulted for each of the three writing genres, it still may be beneficial for teachers to provide explicit instruction on specific vocabulary associated with each genre of writing. However, there is overlap across many of the scoring elements no matter which genre is taught because as children learn to develop an organized and on-topic writing piece, they make improvements in other skills (e.g., grammar and vocabulary; McKenna et al., 2021).

Interrater Reliability

When analyzing the interrater reliability scores there was great variability. A small percentage of teachers achieved exact agreement at or above 80% for the entire rubric and for scoring elements when their scores were compared with the scores of the researcher. Punctuation was the scoring element with the highest percentage of exact interrater agreement, 36% of the 56 rating teams achieved 80% or higher; likewise, this was shown in the highest agreement percentage of rating trios, 56% of the 43 trios, which also

was reflected in a Krippendorff's α above .80. When looking at each pair of raters, 73% of the 56 rating teams had an exact agreement of 80% or higher on one or more of the 10 scoring elements. Although the only modest agreement was achieved when assessing for exact agreement, a high degree of concordance was attained when the tolerance of scores was widened to within 1 point. The averages of the 1-point discrepancy agreement scores were much more promising. The average for the rubric scores within 1-point averaged 91% agreement. Letter formation had the highest average of ratings within 1 point at 96% and spelling had the lowest average at 81% agreement. In this case, scoring would be considered in the acceptable agreement range for all rubric elements. When reviewing data from each trio, 88% of the 43 trios obtained an α above .80 at least 1 time.

Reliability data indicate that future efforts are required to improve the interrater agreement, such as further revisions to the rubric, the scoring manual, and training opportunities made available to educators. The reliability data obtained were higher than data from previous studies when teachers scored writing (James-Burdumy et al., 2009). Burgin and Hughes (2009) commented on the subjectivity of scoring writing even when teachers worked in teams to score.

Participants' feedback will guide revisions. Teachers often indicated that they never had college courses that addressed writing. Clear expectations for scoring need to be established. At times teachers crossed out a criterion on a specific scoring box to give a higher score. There is a need for training to allow teachers to fully understand each scoring element. More guidance is needed on how to use objective data to assign scores. The revisions to the wording, training opportunities, and additions of expectations for teachers are likely to increase reliability. The pretesting procedures also can be changed for future iterations. For example, participants could complete pretesting individually instead of in a group or while receiving training on the rubric. This would allow the researcher to collect data from participants via a survey system, analyze data, provide feedback to teachers, and resolve reliability issues prior to scoring the assessment.

Participant Input on Social Validity

An innovative feature of this study was the involvement of educators in each stage of the development. Existing literature often reports on graduate students serving as coders (e.g., Kim et al., 2015; Sturm et al., 2012), teacher teams working together on scoring (e.g., Burgin & Hughes, 2009), or teachers scoring using a measure created or adopted by a district (e.g., Gansle et al., 2006). These studies used scoring rubrics that were generated or adapted by researchers or the district. Only one of these studies solicited feedback from a small number of stakeholders ($n = 4$) after they used the measure. Opportunities for the teachers

and SLPs to identify potential improvements to consider were absent from prior studies (e.g., Sturm et al., 2012). Translating research into practice becomes challenging when stakeholders who are expected to use the measure have minimal involvement or opportunities to provide input during development.

Revisions may require a close review of the criteria in each box. The spelling scale was originally generated using a K and first-grade phonics continuum. However, spelling in the writing samples collected did not always follow this progression. For example, students consistently used long vowel patterns (e.g., ea and ai), blends, and diagraphs, whereas CVCe words were absent or misspelled. Teachers found the vocabulary scale challenging to use. Teachers wondered if it would be possible to relate the word choice in the writing samples to the purpose of the writing prompt. Many teachers indicated that they taught vocabulary differently; thus, it became difficult to assess the writing of students who were not in their classroom. In addition, teachers reported that the counts on the rubric were distracting and that there would never be enough time to use these, especially when scoring the writing for every student in their classroom. Rewording is necessary to ensure the descriptions are perceived as less overwhelming to stakeholders using the rubric.

Limitations

Despite the large number of writing samples gathered for this study and the diversity of participants, there were barriers to recruitment. None of the classrooms had 100% student participation, and the rate of consent form return in the Title I schools was lower than non-Title I schools. Because writing is part of everyday educational practice and the exempt status of this research, it is possible that other districts and institutional review boards might grant a waiver of parental permission for students contributing deidentified writing samples for measurement development. In this study, data were only collected once, in January. In the future, data need to be collected multiple times during a school year. The teachers were not blinded to the students in their own classrooms, which may have biased their scoring. In addition, the writing samples were not typed out, so handwriting may have influenced scoring. This study took place in one large district and was developed with teachers all using the same state standards, curriculum, and instructional resources. The standards implemented in Florida at the time of this study were the English Language Arts Florida Standards (LAFS), a revised version of the Common Core State Standards (CCSS). However, there is sufficient overlap of the writing elements addressed in the LAFS and CCSS and these standards share common objectives. Further exploration of generalizability to teachers across the nation is needed.

Future Research

The preliminary data collected were analyzed descriptively to assess differences in kindergarten and first-grade writing. Although a large number of writing samples were collected from students across a variety of subgroups and levels of ability, future research needs to be conducted to establish norms using stratified random sampling. Additional studies need to further analyze differences between grade levels and student demographic characteristics (e.g., race, ethnicity, and gender). Future data collection could be used to derive benchmarks for beginning, middle, and end of kindergarten and first grade. Studies of how student writing correlates to reading performance and report card grades in English language arts are needed. Future research may derive a single score for writing proficiency empirically. This will allow for a quantitative value to represent writing during quarterly report card reviews and assist teachers as they grade students. Educators who participate in future studies will be asked to report the time it takes to score the writing samples. Analyses will be conducted to determine if this time decreases after opportunities to score writing samples collected over multiple assessment periods.

Conclusion

This project holds promise for offering an efficient and effective measure for assessing early writing. Kindergarten and first-grade teachers may have access to a comprehensive rubric that can inform data-driven decision-making and clearly communicate expectations. Educators may use the rubric to plan core (Tier 1) instruction and to identify instructional areas of focus for students needing Tier 2 writing intervention. The rubric is designed so that educators may choose to use the entire assessment 3 to 4 times per year. If they are looking to assess the specific skills addressed during the whole group and/or small group instruction, they may choose to use elements of the rubric. For example, to assess sentence writing, they can score grammar, variety, upper/lowercase letter use, and punctuation. Accountability for assessment of writing in kindergarten and first grade may increase if teachers have access to a universal measure that can be used with relative ease. Continued iterative development of the EEWR has the potential to make a meaningful contribution to the writing assessment of young children.

Acknowledgments

Thank you to the district leaders, principals, assistant principals, teachers, coaches, and speech-language pathologists who supported and participated in this project. We have learned so much and are grateful for the perspective you share to ensure that research is applicable to everyday practice. This project was only possible because of your partnership!



Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

Grant R324B180004 from the Institute of Education Sciences, U.S. Department of Education, awarded to the University of Kansas supported the time spent making minor revisions to this article. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

ORCID iDs

Meaghan McKenna  <https://orcid.org/0000-0001-9430-415X>
Howard Goldstein  <https://orcid.org/0000-0003-3270-2739>

Supplemental Material

Supplemental material for this article is available on the *Assessment for Effective Intervention* website with the online version of this article.

References

- Allen, A. A., Jung, P. G., Poch, A. L., Brandes, D., Shin, J., Lembke, E. S., & McMaster, K. L. (2020). Technical adequacy of curriculum-based measures in writing in grades 1–3. *Reading & Writing Quarterly, 36*(6), 563–587. <https://doi.org/10.1080/10573569.2019.1689211>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*.
- Applebee, A. (2000). Alternative models of writing development. *Perspectives on Writing Research, Theory, and Practice, 90*–110.
- Beck, I. L., McKeown, M. G., & Kucan, L. (2013). *Bringing words to life: Robust vocabulary instruction*. Guilford Press.
- Bondie, R. S., Dahnke, C., & Zusho, A. (2019). How does changing “one-size-fits-all” to differentiated instruction affect teaching? *Review of Research in Education, 43*(1), 336–362. <https://doi.org/10.3102/0091732X18821130>
- Burgin, J., & Hughes, G. D. (2009). Credibly assessing reading and writing abilities for both elementary student and program assessment. *Assessing Writing, 14*(1), 25–37. <https://doi.org/10.1016/j.asw.2008.12.001>
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Routledge Academic.
- Coker, D. L., & Ritchey, K. D. (2010). Curriculum-based measurement of writing in kindergarten and first grade: An investigation of production and qualitative scores. *Exceptional Children, 76*(2), 175–193.
- Coker, D. L., & Ritchey, K. D. (2014). Universal screening for writing risk in kindergarten. *Assessment for Effective Intervention, 39*(4), 245–256. <https://doi.org/10.1177/1534508413502389>

- Conway, J. M., Jako, R. A., & Goodman, D. F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology, 80*(5), 565–579.
- Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart, and Winston.
- Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1*(1), 16–29. <https://doi.org/10.1037/1082-989X.1.1.16>
- Espin, C. A., Weissenburger, J. W., & Benson, B. J. (2004). Assessing the writing performance of students in special education. *Exceptionality, 12*(1), 55–66.
- Feenstra, H. M. (2014). *Assessing writing ability in primary education: On the evaluation of text quality and text complexity* [Doctoral dissertation]. <https://research.utwente.nl/en/publications/assessing-writing-ability-in-primary-education-on-the-evaluation>
- Frey, B. B. (Ed.). (2018). *The SAGE encyclopedia of educational research, measurement, and evaluation*. Sage Publications.
- Gansle, K. A., VanDerHeyden, A. M., Noell, G. H., Resetar, J. L., & Williams, K. L. (2006). The technical adequacy of curriculum-based and rating-based measures of written expression for elementary school students. *School Psychology Review, 35*(3), 435–450.
- Hall-Mills, S., & Apel, K. (2015). Linguistic feature development across grades and genre in elementary writing. *Language, Speech, and Hearing Services in Schools, 46*(3), 242–255. https://doi.org/10.1044/2015_LSHSS-14-0043
- Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1*(1), 77–89.
- Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement, 66*(3), 393–416.
- Huot, B. (2002). Toward a new discourse of assessment for the college writing classroom. *College English, 65*(2), 163–180.
- James-Burdumy, S., Mansfield, W., Deke, J., Carey, N., Lugo-Gil, J., Hershey, A., . . . Faddis, B. (2009). *Effectiveness of selected supplemental reading comprehension interventions: Impacts on a first cohort of fifth-grade students*. NCEE 2009-4032. National Center for Education Evaluation and Regional Assistance.
- Jeffery, J. V. (2011). Subjectivity, intentionality, and manufactured moves: Teachers' perceptions of voice in the evaluation of secondary students' writing. *Research in the Teaching of English, 46*(1), 92–127.
- Jeong, H. (2017). Narrative and expository genre effects on students, raters, and performance criteria. *Assessing Writing, 31*, 113–125. <https://doi.org/10.1016/j.asw.2016.08.006>
- Juel, C., Griffith, P. L., & Gough, P. B. (1986). Acquisition of literacy: A longitudinal study of children in first and second grade. *Journal of Educational Psychology, 78*(4), 243–255.
- Kaminski, R. A., & Good, R. H., III. (1998). Assessing early literacy skills in a problem-solving model: Dynamic indicators of basic early literacy skills. In M. R. Shinn (Ed.), *The Guilford school practitioner series. Advanced applications of curriculum-based measurement* (pp. 113–142). Guilford Press.
- Keller-Margulis, M. A., Ochs, S., Reid, E. K., Faith, E. L., & Schanding, G. T., Jr. (2019). Validity and diagnostic accuracy of early written expression screeners in kindergarten. *Journal of Psychoeducational Assessment, 37*(5), 1–14. <https://doi.org/10.1177/0734282918769978>
- Kim, Y. S., Al Otaiba, S., Folsom, J. S., Greulich, L., & Puranik, C. (2014). Evaluating the dimensionality of first-grade written composition. *Journal of Speech, Language, and Hearing Research, 57*(1), 199–211. [https://doi.org/10.1044/1092-4388\(2013\)12-0152](https://doi.org/10.1044/1092-4388(2013)12-0152)
- Kim, Y. S., Al Otaiba, S., Puranik, C., Folsom, J. S., Greulich, L., & Wagner, R. K. (2011). Componential skills of beginning writing: An exploratory study. *Learning and Individual Differences, 21*(5), 517–525. <https://doi.org/10.1016/j.lindif.2011.06.004>
- Kim, Y. S., Al Otaiba, S., Wanzek, J., & Gatlin, B. (2015). Toward an understanding of dimensions, predictors, and the gender gap in written composition. *Journal of Educational Psychology, 107*(1), 79–95. <https://doi.org/10.1037/a0037210>
- Krueger, R. A. (1994). *Focus groups: A practical guide for applied research* (2nd ed.). SAGE.
- Lam, R. (2018). *Portfolio assessment for the teaching and learning of writing*. Springer.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.
- Lembke, E. S., Garman, C., Deno, S. L., & Stecker, P. M. (2010). One elementary school's implementation of response to intervention (RTI). *Reading & Writing Quarterly, 26*(4), 361–373. <https://doi.org/10.1080/10573569.2010.500266>
- Lloyd-Jones, R. (1977). Primary trait scoring. *Evaluating Writing: Describing, Measuring, Judging, 33–66*.
- Lonigan, C. J., & Milburn, T. F. (2017). Identifying the dimensionality of oral language skills of children with typical development in preschool through fifth grade. *Journal of Speech, Language, and Hearing Research, 60*(8), 2185–2198. https://doi.org/10.1044/2017_JSLHR-L-15-0402
- Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research, 35*, 382–385.
- MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*(1), 84–99.
- McCoach, D. B., Gable, R. K., & Madura, J. P. (2013). *Instrument development in the affective domain*. Springer. https://doi.org/10.1007/978-1-4614-7135-6_8
- McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review, 15*(1), 28–50. <https://doi.org/10.1177/1088868310366253>
- McKenna, M., Goldstein, H., Soto, X., Cheng, K., Troia, G., & Ferron, J. (2021). Supplemental intervention improves writing performance of first grade students: A single case experimental design evaluation. *Journal of Educational Research, 114*(3), 278–293. <https://doi.org/10.1080/00220671.2021.1923450>
- McMaster, K. L., & Espin, C. (2007). Technical features of curriculum-based measurement in writing: A literature review. *The Journal of Special Education, 41*(2), 68–84.
- McMaster, K. L., Shin, J., Espin, C. A., Jung, P. G., Wayman, M. M., & Deno, S. L. (2017). Monitoring elementary students'

- writing progress using curriculum-based measures: Grade and gender differences. *Reading and Writing*, 30, 2069–2091. <https://doi.org/10.1007/s11145-017-9766-9>
- Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 30(10), 955–966.
- Muthen, L. K., & Muthen, B. O. (2019). *Mplus [computer software]*. Muthén & Muthén.
- National Center for Education Statistics. (2012). *The nation's report card: Writing 2011*. Institute of Education Sciences, U.S. Department of Education.
- National Commission on Writing. (2006). *Writing and school reform*. <http://www.collegeboard.com>
- Northwest Regional Educational Laboratory 6+1 Trait® Writing. (2014). <http://www.nea.org/tools/lessons/59760.htm>
- Olinghouse, N. G. (2008). Student-and instruction-level predictors of narrative writing in third-grade students. *Reading and Writing*, 21(1–2), 3–26.
- Payan, A. M., Keller-Margulis, M., Burrige, A. B., McQuillin, S. D., & Hassett, K. S. (2019). Assessing teacher usability of written expression curriculum-based measurement. *Assessment for Effective Intervention*, 45(1), 51–64. <https://doi.org/10.1177/1534508418781007>
- Philippakos, Z. A. T., & FitzPatrick, E. (2018). A proposed tiered model of assessment in writing Instruction: Supporting all student-writers. *Insights into Learning Disabilities*, 15(2), 149–173.
- Puranik, C., Duncan, M., Li, H., & Ying, G. (2020). Exploring the dimensionality of kindergarten written composition. *Reading and Writing*, 33, 2481–2510. <https://doi.org/10.1007/s11145-020-10053-1>
- Rios, J., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema*, 26(1), 108–116. <https://doi.org/10.7334/psicothema2013.260>
- Romig, J. E., Therrien, W. J., & Lloyd, J. W. (2017). Meta-analysis of criterion validity for curriculum-based measurement in written language. *The Journal of Special Education*, 51(2), 72–82. <https://doi.org/10.1177/0022466916670637>
- Sturm, J. M., Cali, K., Nelson, N. W., & Staskowski, M. (2012). The Developmental Writing Scale: A new progress monitoring tool for beginning writers. *Topics in Language Disorders*, 32(4), 297–318. <https://doi.org/10.1097/TLD.0b013e318272159e>
- Taylor, R. L. (2006). *Assessment of exceptional students: Educational and psychological procedures* (7th ed.). Pearson Education.
- United Kingdom Department for Education. (2012). *The research evidence on writing*. Department for Education. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/183399/DFE-RR238.pdf
- Wagner, R. K., Puranik, C. S., Foorman, B., Foster, E., Tschinkel, E., & Kantor, P. T. (2011). Modeling the development of written language. *Reading and Writing*, 24, 203–220. <https://doi.org/10.1007/s11145-010-9266-7>
- White, K. M. (2013). Associations between teacher–child relationships and children’s writing in kindergarten and first grade. *Early Childhood Research Quarterly*, 28(1), 166–176. <https://doi.org/10.1016/j.ecresq.2012.05.004>
- Willis, G. B. (2004). Cognitive interviewing revisited: A useful technique, in theory. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singe (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 23–43). Wiley.