Expanding Assessment to Instructionally Relevant Writing Components in Middle School

Adrea J. Truckenmiller, Eunsoo Cho, and Gary A. Troia

Michigan State University

August 13, 2022

Author Note: This research was supported in part by a grant from the Institute for Research on Teaching and Learning at Michigan State University and grant #R305A210061 from the U.S. Department of Education, Institute of Education Sciences to the first author and grant #R305A160049 from the U.S. Department of Education, Institute of Education Sciences, to the third author. The authors would like to acknowledge Greenflux, LLC for programming the Writing Architect.

Publication History: Received 23 April 2020; Received in revised form 15 February 2022; Accepted 11 July 2022; Available online 13 August 2022

Citation: Truckenmiller, A. J., Cho, E., & Troia, G. A. (2022). Expanding assessment to instructionally relevant writing components in middle school. *Journal of School Psychology*, *94*, 28–48.https://doi.org/10.1016/j.jsp.2022.07.002

Abstract

Although educators frequently use assessment to identify who needs supplemental instruction and if that instruction is working, there is a lack of research investigating assessment that informs what instruction students need. The purpose of the current study was to determine if a brief (approximately 20 min) task that reflects a common middle school expectation (writing in response to text) provides educators with information about students' strengths and weaknesses in four research-based components of writing. Results indicated that, at the end of elementary school (Grade 5), students' word- and sentence-level errors, text-level plan, and typing fluency predicted 43% of their performance in written composition quality and all these factors play a role in writing achievement. At the end of middle school (Grade 8), text-level plan and word-level accuracy remained important components. Implications for using assessment to guide selection of evidence-based writing instruction throughout middle school are discussed.

Keywords: Writing assessment, Middle school writing, Data-based decision making

Students vary widely in their written composition performance, with most students (72%–74%) performing below proficiency expectations throughout elementary, middle, and high school (National Assessment of Educational Progress (NAEP), U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 2011). Research consistently demonstrates heterogeneity in writing skills within a single grade (i.e., large standard deviations), which implies students have great variability in their instructional needs (e.g., Dockrell et al., 2018). Given the wide variability of students' writing performance, schools need assessment tools to differentiate the instructional needs of the majority of their students by using measures that help educators identify appropriate instructional targets (Berninger & O'Malley May, 2011; Berninger & Wolf, 2009).

Diagnostic assessments are a broad category of assessments that can differentiate instruction by identifying the specific skills for which individual students need assistance. Diagnostic assessment information can guide selection of instructional targets that have an impact on proximal and distal writing outcomes for students who struggle with one or more aspects of writing. For example, if a student scores below benchmark on a written composition achievement task (e.g., state test) or a general outcome task (e.g., curriculum-based measurement in written expression), additional diagnostic assessment may reveal that the student needs more skill development in spelling and text organization. Through the provision of targeted evidence-based instruction in the component skills of spelling and text organization, one would hypothesize that the student would practice and integrate those component skills in their compositions and thus improve in overall writing performance as evaluated through general outcome and achievement assessment (Kent & Wanzek, 2016; Torrance et al., 2020). Currently, this level of diagnostic detail requires extensive assessment resources (e.g., administering and scoring norm-referenced diagnostic assessments of written expression) that are prohibitive in a typical classroom context, where >70% of the students are struggling with written composition.

A classroom-based diagnostic assessment that provides details about specific strengths and weaknesses to inform which skills or strategies should be emphasized in instruction would be informative for teachers, school psychologists, interventionists, and problem-solving team members (Berninger & O'Malley May, 2011; Graham et al., 2016). Although researchers and practitioners acknowledge a need for more useable writing assessments, it has been difficult to achieve a balance of breadth, depth, efficiency, and technical adequacy in writing assessment for such a large range of writing performance (Institute for Education Sciences, Technical Working Group, 2017). The current study explores the development of a new diagnostic assessment, the Writing Architect (WA), that was designed, in part, to connect writing skill performance with instructional decisions.

Writing Architect

The WA contains a set of written composition prompts administered via the web and human-scored for a variety of potential instructional targets. The current study focused on the informational writing prompts for Grades 5 and 8. These prompts include a grade-level appropriate informational text that the student listens to (and reads) and a text-dependent question that elicits a written composition. The students are given 3 min to plan their response (using blank paper and a pencil) and 15 min to compose their full response, typed on the web-based application. See Truckenmiller et al. (2020) for empirical evidence that the performance on this task (scored with a general outcome metric and a writing quality rubric) predicts written composition outcomes. Although the general outcome metric demonstrated utility for monitoring progress and the quality rubric could identify *who* needs supplemental instruction, a writing expert would be needed to interpret *what* instruction those students would require. Following others who have suggested that component skills of writing may be useful for assessment and instruction (Berninger & Wolf, 2009; Kent & Wanzek, 2016; Limpo et al., 2017), we sought evidence of validity for four component skills measured with the WA. These four scores include (a) typing fluency, (b) word-level errors, (c) sentence-level errors, and (d) text-level plan. The goal for the WA is for teachers to have a tool to examine the wide variability in writing skills in their classrooms (Dockrell et al., 2018) that will provide information about malleable component skills of writing (Limpo et al., 2017) that are impacted by research-based instructional practices (Graham et al., 2016; Santangelo & Olinghouse, 2009). This complex interaction of assessment, students' component skill performance, and teachers' interpretation of both the assessment results and students' skill development requires careful attention to theories of assessment development, writing development, and multiple pieces of evidence for validity. We integrated each of these considerations in the development of the WA. We use the National Research Council (NRC; 2001) assessment development framework, the Assessment Triangle, to contextualize the development of the WA.

Diagnostic Assessment Development

In their seminal work describing the Assessment Triangle, the NRC places *observation*, *cognition*, and *interpretation* at the vertices of the triangle and describe the connection between each of the three vertices (NRC, 2001). For educators to use assessment results to inform instruction, the assessment task must (a) reflect a valued performance outcome (i.e., observation), (b) measure domains represented in theories or statistical models (i.e., cognition), and (c) have evidence for interpretation that leads intuitively to instructional decisions (i.e., interpretation). In the WA, we anchor observation to current shared values/standards for writing performance. For cognition, we evaluate four constructs that represent some of the key components of the Simple View of Writing (Berninger & Winn, 2006) and the Direct and Indirect Effects of Writing model (Kim, 2020; Kim & Schatschneider, 2017). Finally, for interpretation, we use Messick's (1995, 1998) seminal framework to specify multiple sources of construct validity evidence needed for improving use of classroom written composition assessment.

Observation

Educationally relevant assessment must allow educators to *observe* student performance that is representative of valued outcomes (NRC, 2001) and the socio-cultural role of the written activity (Graham, 2018). Written composition expectations and the assessment of those expectations have been constantly evolving (Mo & Troia, 2017). In the intersection of modern expectations for written composition and a lack of available classroom assessments are three features that need to be considered for observation. These features include (a) writing in the informational genre, (b) using evidence from a text in the written response, and (c) typing the composition. A writing task that incorporates these three features would be valuable not only in language arts classes, but also social studies and science, as well as by colleges and universities, employers, and others in society where social and civic engagement are predicated on using evidence from text and communicating high quality information both online and offline (National Commission on Writing, 2004; National Governors Association and Council of Chief School Officers, 2010). Writing to learn social studies and science content knowledge is particularly effective for social studies and science achievement (Bangert-Drowns et al., 2004; Graham et al., 2020). Furthermore, everyone uses informational writing on an electronic device in their daily civic life. Examples include writing directions about performing a task at work, emailing project details to a colleague, texting a family member medical or public health information, and writing a social media post about a current news story.

Accordingly, digital typing of informational text is intentionally included in writing proficiency assessment, including the most recent iterations of the National Assessment of Education Progress (NAEP) and most state tests (e.g., Partnership for Assessment of Readiness for College and Careers, and Smarter Balanced Assessment Consortium [SBAC]).

In the current study, the Michigan Student Test of Educational Progress (MSTEP) was used as the primary distal outcome. The MSTEP has an extended written composition task (that could be informational, persuasive, or narrative) in Grades 5 and 8 and uses items from the SBAC. Extended written composition tasks on state tests are typically only administered once in the elementary grades (Grade 4 or 5), once in middle school (Grade 8), and once in high school (Grade 10, 11, or 12). Therefore, information about students' written composition achievement is available to teachers infrequently. Perhaps the infrequency is due to the time-consuming nature of assessing writing with this kind of task or maybe due to the fact that average writing performance does not change much across grade levels (e.g., see scaled scores on the NAEP). Regardless of the reason, educators need more frequent access to observation of what their students write to guide writing instruction in the classroom. Therefore, they must have a proximal observation that relates to the distal state outcome. Many educators use released items from state tests or create their own prompts. To score written compositions, schools adopt and adapt the rubrics from the state test as their proximal measure and many teachers cite the rubric as their guide for writing instruction (Applebee

& Langer, 2011). Although rubrics across schools differ, most rubrics of writing quality include considerations about writing purpose, supporting details, introduction, conclusion, organization, coherence, cohesion, language use, and mechanics (Troia et al., 2019).

The rubric proximal outcome is very useful for various purposes in the classroom, but not all purposes. Using rubrics to tailor individualized feedback to students has been identified as a best practice for improving writing achievement as well as guiding feedback on the process of writing (i.e., planning, composing, and revising; Graham et al., 2016). Teachers who have training with self- regulated strategy instruction also use the rubric as an effective method for choosing which writing process strategies to teach to their class (Rouse & Kiuhara, 2017). Therefore, writing quality rubrics provide information for individualized feedback on the writing process and strategy instruction (Graham et al., 2016).

Although the rubric provides educators with useful information about providing strategy instruction and process instruction, teachers report that it does not provide them with more fine-grained information about specific writing *skills* that students need (McKeown et al., 2019). Research also suggests that rubric scores reflect a unidimensional construct that cannot validly differentiate performance with specific components of writing because rated performance on one dimension of quality on a rubric heavily influences rated performance on the other dimensions (Gansle et al., 2006; Kim et al., 2015; Lee et al., 2008; Troia et al., 2013). This suggests that more fine-grained information is needed to reflect specific skill components of writing. Researchers with an interest in promoting strong connections between research and practice suggest the need for instruments that clearly link the components underlying writing performance with specific instructional actions for addressing writing needs (Berninger & O'Malley May, 2011; McCardle et al., 2018).

Cognition: Research-Based Components of Writing

The cognition vertex of the Assessment Triangle is defined by the cognitive constructs that are represented in theory and established in empirical models (NRC, 2001). To connect cognition with observation in the Assessment Triangle, the constructs measured by the assessment must demonstrate an empirical relation to the observation (i.e., the proximal quality rubric and distal achievement score). Many theoretical and empirical models of written composition exist (see O'Rourke et al., 2018, for a comprehensive review). Most models combine a variety of malleable component skills, developmental factors, and executive functioning abilities. To find the component skills that are most malleable, we looked at a meta-analysis of writing skills conducted by Kent and Wanzek (2016). They found the highest effect sizes for reading (ES = 0.48), spelling (ES = 0.44), handwriting fluency (ES = 0.34), and oral language (i.e., vocabulary and grammar; ES = 0.32). In the present study, we explored each of these areas except reading. We did not include a measure of reading because most schools already administer separate measures of reading and because a large amount of variance is shared between reading and writing (Abbott et al., 2010; Truckenmiller & Petscher, 2020).

Except for reading, the malleable components found by Kent and Wanzek (2016) generally align with the Simple View of Writing (Berninger & Winn, 2006). This theory posits that text generation at the word-, sentence-, and discourse-levels are supported by transcription (handwriting, typing, and spelling) and executive functions/self-regulation (i.e., planning, reviewing, and revising; Berninger et al., 2002). Berninger et al. (2002) and Berninger and Wolf (2009) proposed that combinations of these components may be impaired and interventions with demonstrated effects on these components be implemented to improve writing outcomes. However, this may be more easily said than done in writing. Unfortunately, writing assessment, writing theory, and writing instruction do not intuitively align well given significant overlap of the component constructs and limitations in available assessments (McCardle et al., 2018).

The complex interrelations between component constructs are highlighted in a few recent studies that are beginning to evaluate the dimensionality of components of writing. In early elementary school, Kim and Schatschneider (2017) provided strong empirical evidence for the Direct and Indirect Effects model of Writing (DIEW) that has similar components as the SVW. In the DIEW model, working memory exerted a direct effect on writing and an indirect effect through each of the components of writing they measured. Handwriting, spelling, and discourse-level text generation also had direct effects on written composition. The discourse-level ideation component was a second-order factor comprised of vocabulary and grammar (i.e., sentence construction), which was mediated by inferencing, theory of mind, and comprehension monitoring. When the DIEW model was tested in Grade 4, Kim (2020) found support for the same constructs with a larger contribution by language skills. The representation of text generation in DIEW as word-, sentence-, and discourse-level language skills is consistent with multiple studies of writing in late elementary and middle school grade levels (Abbott et al., 2010; Berninger et al., 1994; Dockrell et al., 2018; Kim et al., 2015).

In later middle school, Limpo et al. (2017) found that handwriting fluency, spelling accuracy, translating skills (i.e., sentence combining and syntactic correctness), and discourse-level planning were all correlated components that together predicted 43% of the variance in writing quality in the opinion genre, with each component making a unique contribution. They found that the influence of handwriting fluency on writing quality was mediated by discourse-level planning and that the effect of spelling accuracy on quality was mediated by translating. Regardless of the interrelations of the components, studies of middle school written composition have evaluated the same four components as those included in the study by Limpo and colleagues and demonstrated some type of unique contribution for each component (e.g., Bereiter & Scardamalia, 1987; Fitzgerald & Markham, 1987; Graham et al., 1995; Koutsoftas, 2016, 2018; Lienemann et al., 2006; Saddler & Asaro, 2007; Tracy et al., 2009; Troia et al., 2019; Troia & Graham, 2002; Wagner et al., 2011).

In the present study, we aimed to balance theoretical construct coverage, empirical linkages to proximal and distal outcomes, and connections to evidence-based intervention by measuring four of the most prominent components of the above models. We included (a) word-level errors, (b) sentence-level errors, (c) text-level plan, and (d) typing fluency. In the sections that follow, we examine each of these four components more closely to show how they connected each vertex of the Assessment Triangle. We demonstrate each component's role in writing development (i.e., cognition vertex), how they are typically scored (i.e., observation vertex), and the interventions that have had an impact on each component (i.e., interpretation vertex).

Typing Fluency. In a writing task that requires composing on the computer, typing is the required mode of transcription. Typing and handwriting both involve coordinated motor movements to produce orthographical representations of the morphophonemic spelling of words in the English language (Berninger et al., 1994). Typing and handwriting also are different in that handwriting provides additional language input with the formation of letters (Troia et al., 2020). Therefore, we anticipated that typing fluency would play a similar role, but likely slightly diminished, as handwriting fluency on writing outcomes.

Handwriting consistently plays a constraining role on writing quality, from the early elementary grades through early middle school, when it is measured in conjunction with spelling (Bourdin & Fayol, 2002; Feng et al., 2019; Graham et al., 1997; Jones & Christensen, 1999; Kent & Wanzek, 2016; Puranik & Al Otaiba, 2012; Wagner et al., 2011) and when it is separated from spelling abilities (Christensen, 2005; Connelly et al., 2007). However, in later grades (i.e., Grades 7 and 8), transcription is more automatized and exerts an indirect effect on writing outcomes. For example, Limpo et al. (2017) found that the role of handwriting fluency was fully mediated by planning skills to predict writing outcomes. This corresponds with cognitive models demonstrating that when first learning to handwrite or type, students must devote cognitive resources to the motor planning for selecting and producing the correct orthographic form. It takes time and practice to reach a level of automaticity that frees those cognitive resources for text generation and other higher-order processes (Olive, 2014).

Although both are considered transcription skills, handwriting and spelling tap into different processes (Limpo et al., 2017). Therefore, intervention requires explicit attention to the motor movements for handwriting/typing and to the phonemic, morphologic, and orthographic features of language for spelling. A meta-analysis of explicit handwriting instruction indicated large gains in writing quality (ES = 0.84) and quantity (ES = 1.33; Santangelo & Graham, 2016). Surprisingly, these effects were not statistically different between Grades 1–4 and Grades 5–9, indicating that handwriting may continue to be challenging for some writers in later grades and associated remediation is effective. Fewer studies have focused specifically on typing, but a meta-analysis by Goldberg et al. (2003) indicated moderate effects of word processing instruction on text quality (ES = 0.41) and quantity (ES = 0.50) for kindergarten through Grade 12 students.

Word-Level Errors. In the present study, word-level errors were defined as the percentage of words misspelled or lacking obligatory capitalization. Spelling accuracy (in isolation or in text) is the most common measurement of spelling in writing studies (e.g., Graham et al., 1997; Limpo et al., 2017; Wagner et al., 2011). Spelling is one of the more robust predictors of writing abilities across a lifetime (e.g., Abbott et al., 2010; Graham et al., 1997; Kim et al., 2015; Olinghouse, 2008). Not only does spelling have a direct impact on written composition quality, it also has an indirect relationship through vocabulary choice (Graham & Santangelo, 2014; McCutchen, 2011). When a student tries to generate vocabulary words for a text, they may forego more precise terms for words they know how to easily spell. For example, in informational writing, a student may be more likely to write the word 'idea' instead of a more precise academic word like 'hypothesis'. In the current study, we used the term word-level errors (Berninger et al., 1994) instead of spelling accuracy. We use the term word-level to examine both the accuracy (spelling) and vocabulary choice at the word level. We examined the relation of spelling to vocabulary use to see if word-level errors as a measure tapped not only transcription, but also the word-level portion (i.e., vocabulary) of text generation.

Spelling instruction can have a moderate impact on students' spelling abilities within writing in Grades 3–6 (ES = 0.62) and a smaller, but significant impact in Grades 7–12 (ES = 0.31; Graham & Santangelo, 2014). Spelling instruction that is explicit and sequenced has a larger impact on students' spelling skills than implicit instruction (ES = 0.94), and this finding is independent of grade level and other literacy abilities; additionally, the gains in spelling are maintained over time (ES = 0.53). It is important to note that, within Graham and Santangelo's meta-analysis, the indirect effects of spelling instruction on writing quality and quantity were not significant, suggesting that spelling may be just one component of effective written composition instruction (Berninger et al., 2002).

Sentence-Level Errors. Sentence-level errors include missing words, missing or incorrect punctuation (e.g., run-on sentences), subject-verb disagreements, noun-modifier disagreements, and other blatant syntactic errors (e.g., words out of order). These errors are typically captured in curriculum-based measurement tasks as incorrect writing sequences, along with spelling and capitalization errors. For the purposes of identifying potentially different instructional implications in this study, we separated the sentence-level errors in incorrect writing sequences from the word-level errors (spelling and capitalization). Errors at the sentence-level can represent constraints in the *translation* process of oral language to written language (Limpo et al., 2017), as well as individual variation in sentence-level language ideation (e.g., Koutsoftas & Gray, 2012; Koutsoftas & Petersen, 2017). The production of written words in syntactically correct and increasingly complex sentences builds upon transcription processes by also including ordering of words, referencing previous words (e.g., anaphora, connective words and phrases), and other conventions of the English sentence (Berman & Verhoeven, 2002; Limpo et al., 2017). When included with other aspects of writing, sentence-level translating predicts significant variance in writing outcomes (Beers & Nagy, 2011; Limpo, Alves and Connelly, 2017;

Truckenmiller & Petscher, 2020). Moreover, correcting sentence level errors often is required as part of state achievement tests of writing (e.g., SBAC).

Sentence-level translation variability signals students' capacity to create coherent and logical sentences and, when students are developing this aspect of their writing, they have benefitted from instruction where they learn to combine shorter simple sentences to form more complex and grammatically correct sentences (for reviews, see Datchuk & Kubina, 2012, and Graham & Perin, 2007). A meta-analysis demonstrated moderate effects (ES = 0.50) for sentence combining instruction in grades 4 through 9 (Graham & Perin, 2007). Therefore, we expected that our sentence-level accuracy metric would be most related to students' performance on a sentence- level conventions assessment (i.e., punctuation) and sentence combining.

Text-Level Plan. In the WA, students are given a blank sheet of paper before they type their composition and are prompted to plan their composition using methods they have learned in school. The act of planning draws on multiple aspects of writing, including writing process, self-regulation, working memory, and discourse structure (De La Paz, 2007). In cognitive models of writing, planning refers to generation and organization of ideas (Kellogg et al., 2013). In other words, skilled writers use their knowledge of the discourse structure of text given the purpose for which they are writing to help them plan (e.g., McCutchen, 2011). In the present study, we evaluated the information on each student's planning sheet to provide insight into the student's conceptualization of the organization of the text as a whole (i.e., discourse-level text generation) using a 4-point scale (described in the Method section). The scale reflects the progression of students from a “knowledge-telling” approach (Bereiter & Scardamalia, 1987), which refers to simply listing ideas without attention to larger organizational schemes, an approach used by many students throughout elementary and middle school, to a deliberate organizational strategy to structure ideas (e.g., a web, headings with bulleted lists), which tends to yield higher quality texts (Englert et al., 1988; Troia et al., 1999). In the context of informative writing, when a student is comparing and contrasting two or more concepts presented in source material for instance, they might organize their ideas with an introduction, ways in which the concepts are similar, ways in which they are different, and a conclusion (Englert et al., 1988; Hebert et al., 2018).

Regardless of the way that planning is conceptualized, planning performance has a large and significant relationship to writing quality (e.g., Koutsoftas, 2016, 2018; Lienemann et al., 2006; Saddler & Asaro, 2007; Tracy et al., 2009; Troia et al., 2019; Troia & Graham, 2002). Given that writing quality is so dependent on organization at the discourse level, planning interventions have the largest effects as compared to the interventions for the other levels of language noted earlier. Meta-analyses have indicated that teaching students to engage in planning activities prior to writing their compositions has a mean effect size of 0.54 and specifically teaching text structure has a mean effect size of 0.59 (Graham et al., 2012).

Interpretation

Improvement of instructional practice is the primary goal of assessment validity (Gersten et al., 1995). Therefore, all assessment development work should lead to a well-defined interpretation. The considerations above for the observation and cognition vertices of the Assessment Triangle were coordinated to support the interpretation vertex. Interpretation is defined as the claims that will be made from the assessment results (NRC, 2001). In his seminal text guiding the field of educational assessment, Messick (1998) described construct validity as the gathering of multiple pieces of evidence to increase confidence in the interpretation of assessment data:

> To validate an action inference requires not only evidence of score meaning but also justification of value implications and action outcomes, especially appraisals of the relevance and utility of the test scores for particular applied purposes and of the social consequences of using the test scores for applied decision making (p. 3).

To meet this goal, Messick (1998) specified that construct validity evidence needs to be collected on "test content, substantive processes, score structure, generalizability, external relationships, and testing consequences" (1998, p. 2). The definition of each of the six sources of construct validity evidence is listed in Table 1. Modern educational assessment design reaffirms the need for evidence that demonstrates the connection between test content, interpretation of the constructs that the score represents, the decisions that educators make, and the intended and unintended consequences of those decisions (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014; Gersten et al., 1995). With the WA, the goal is to connect the four diagnostic component scores (i.e., typing fluency, word-level errors, sentence-level errors, and text-level plan) with educators' decisions choosing instruction that has impacts on those four skills, which subsequently improves socially valued proximal and distal writing outcomes. In Table 1, we detail the forms of evidence that we previously collected (i.e., content validity), evidence reported in the current study (i.e., external, substantive, structural, and generalizability), and evidence that we plan to collect in the future (i.e., consequential validity).

Table 1
Types of validity evidence.

Aspect of construct validity (Messick, 1995, 1998)	Evidence needed (Messick, 1995, 1998)	Operationalization in the current study)
Content Validity The task represents the writing domain.	Expert professional judgment	Experts reviewed passages and questions to judge alignment to informational writing expectations.
External Validity The meaning of the scores is substantiated externally by appraising the degree to which empirical relationships with other measures – or the lack thereof – are consistent with that meaning	Empirical evidence of convergent and discriminant relationships with external measures.	Diagnostic scores are correlated with other gold standard measures of the same construct and are less correlated with measures of related but different constructs.
Structural Validity The internal structure of the assessment should reflect the internal structure of the construct domain.	Scoring should be rationally consistent with the structural components and relate to the construct.	Together, the diagnostic scores make up a significant portion of the variance in writing quality scores.
Substantive Validity The task requires students to use the processes and domain content intended.	Correlation patterns among part scores	Diagnostic scores, when mediated by overall writing quality, predict writing outcomes.
Generalizability Validity The assessment tasks, populations, and/or settings to which the scores apply.	Extent to which score properties and interpretations generalize across different population groups, setting, and tasks.	Evaluated for two populations (Grade 5 and Grade 8).
Consequential Validity Intended and unintended consequences of score interpretation.	Identify sources of potential construct underrepresentation and construct- irrelevant variance. Users' interpretation of scores and use of scores (Gersten et al., 2020)	Included audio reading of the text to reduce impact of reading abilities. Future study expanding construct coverage. Future qualitative study of educator's use of scores.

Purpose of the Present Study

The aim of the present study was to measure four potentially high-leverage writing skills within one relatively brief web-based assessment system in middle school classrooms. Evidence of the *content* validity for the WA task was collected previously. An expert panel consisting of four senior writing researchers reviewed the informational passages and questions and determined that they were aligned to informational writing expectations in the Common Core State Standards (Troia et al., 2020). To further improve content validity across prompts empirically, we also estimated the amount of variance due to the separate prompts, which was 2% or less (Truckenmiller et al., 2020). A previous study also found some evidence of *substantive* validity with the number of correct minus incorrect writing sequences (CIWS) produced in 15 min, mediated by the students' quality of writing measured using a rubric, accounted for 70%–95% of the variance in Grades 5–8 writing achievement (Truckenmiller et al., 2020). Therefore, both the CIWS score and writing quality rubric score were correlated with a valued distal outcome. These scores demonstrated utility for predicting *who* needs additional instruction and may detect if that instruction is working. However, the quality rubric and CIWS do not provide enough information to tell us *why* students struggle with writing. To determine why a student struggles on a classroom writing task, we must measure component writing skills that are most likely to influence student writing achievement (based on empirical models of writing development). Therefore, the present study was conducted to find evidence of external, structural, substantive, and generalizability validity for four component skills of writing, including (a) typing fluency, (b) word-level errors, (c) sentence-level errors, and (d) text-level plan.

Evidence of External Validity

To demonstrate evidence of *external* validity for the four diagnostic scores, we evaluated the correlation of the diagnostic component scores with commonly accepted measures of the same constructs. The National Center for Intensive Intervention (National Center on Intensive Instruction, 2019) identifies a threshold correlation of 0.60 as acceptable evidence of external validity. We hypothesized that word-level errors would be correlated with a standardized norm-referenced measure of spelling near the 0.60 range. Given the theoretical role spelling plays in vocabulary, we also explored the correlation of word-level errors with a standardized norm- referenced assessment of vocabulary to provide a more nuanced understanding of the interrelations of valued and theorized components of writing.

We hypothesized that sentence-level errors would be correlated with a standardized norm-referenced measure of punctuation (one type of sentence-level error) near the 0.60 range. We also expected sentence-level errors to be significantly related to scores on a standardized norm-referenced assessment of sentence combining, which is one proxy for sentence complexity.

Typing fluency and text structure do not have associated standardized norm-referenced tests, but the measures used in the current study reflect a consensus in the field for assessing these components of writing and studies consistently show that they are significantly correlated with the amount and quality of writing students produce.

Evidence of Structural Validity

We hypothesized that the four diagnostic component scores would account for a significant and substantively large amount of variance in classroom writing task performance (as measured by rubric score, the proximal measure of writing performance). Other studies have suggested that typing fluency, word-level errors, sentence-level errors, and text-level plan make up 46% of the variance in Grades 7 and 8 writing performance (Limpo et al., 2017). We acknowledge that there are other components of writing that may increase the variance accounted for and be important for instruction, but those components are already well-represented in the writing quality rubric (e.g., organization, providing supporting evidence for claims).

Evidence of Substantive Validity

We hypothesized that students' performance in the four component skills, as mediated by their writing quality rubric score, would predict their performance on the state writing test. Given that the state test also includes items requiring word- and sentence-level revisions, we hypothesized that there would also be direct relations between word- and sentence-level errors with the state test.

Evidence of Generalizability

We examined how the structural model generalized across two populations that represent important developmental time periods in writing, including at the end of elementary school (Grade 5) and the end of middle school (Grade 8). Understanding the grade levels (or developmental time frames) in which the assessment task is most relevant is an important first step in generalizability before generalization evidence is sought for other groups (e.g., students with disabilities or learning English as a second language).

Method

Participants and Setting

Six general education teachers in three suburban school districts in Michigan volunteered their 15 classrooms to participate in the study. We obtained affirmative parental consent from 75% of the students in the classrooms. We analyzed data from a total of 285 students in Grades 5 (*n* = 175) and 8 (*n* = 110). The demographic data for this convenience sample is provided in Table 2. When compared to a national norm of writing achievement (i.e., the Test of Written Language), this sample performed slightly higher than average with sample mean scores between 0.33 and 1.03 *SD* above the national means. The sample in the present study was the same sample as in the Truckenmiller et al. (2020) study but reports on different scores for different purposes.

Table 2
Demographic information.

	Grade 5 (*n* = 175)	Grade 8 (*n* = 110)
Gender Male		
	90 (51%)	47 (43%)
Female	85 (49%)	63 (57%)
Ethnicity		
Native American		
	2 (1%)	0
Black or African American	11 (6%)	22 (20%)
Hispanic or Latino	11 (6%)	2 (1.8%)
White	107 (61%)	75 (68.2%)
Two or More Races	13 (7%)	3 (2.7%)
Asian	29 (17%)	8 (7.3%)
Economically Disadvantaged	37 (21%)	21 (19%)
Special Education	15 (9%)	6 (6%)
MSTEP Reading, not proficient	30 (17%)	11 (10%)
MSTEP Writing, not proficient	26 (15%)	18 (16%)
English Language Learner	8 (5%)	2 (2%)

Note. Ethnicity, economically disadvantaged, special education, and English language learner information was missing for two participants in Grade 5. MSTEP = Michigan Student Test of Educational Progress.

Materials: The Writing Architect 1.0

The Writing Architect version 1.0 (WA) consists of a web-based application for group administration of writing prompts and a scoring database on the back end. Students are given a writing prompt and asked to (a) plan what they will write (3 min), (b) write their essay (15 min), and (c) test their typing fluency (90 s). In the application, students can listen to the passage prompt (a human voice) and read along on the screen or on a paper copy provided. Next, the prompt specific to the passage appears and students are instructed to spend 3 min planning their response using blank paper provided. Students then are prompted to compose their typed response for 15 min. Students have the option to submit the final response before the end of 15 min. Finally, they are asked to copy a paragraph to measure their typing fluency. The information collected from the web application is transferred to a database for scoring. The database has a customized interface that allows human scoring of the metrics described in the Measures section below.

More details about the students' interaction with the application are provided in the validation study (Truckenmiller et al., 2020).

Informational Passage-Based Prompts

A set of passages were chosen from web services that provide informational articles written for elementary and middle school-aged students. Permission was obtained from the copyright holders to use and modify the passages. A total of five passages for Grade 5 and three passages for Grade 8 were chosen by a researcher (first author) with experience in the development of reading assessments. The word count in the passages ranged from 406 words to 814 words. All passages were professionally judged to represent the informational genre. The informational nature of the passages was confirmed by the Coh-Metrix narrativity score (McNamara et al., 2005), which was below 50% for each passage.

Instructions to the students for each passage were intended to elicit an informational response and required students to use details from the passages to support their answer. For example, students were instructed to "Write an informative paper that will help others learn about building houses out of plastic bottles. Be sure to use information from the article you just read to give reasons why using plastic bottles to build homes would be helpful. Remember, a well written informative paper (1) has a clear main idea and stays on topic, (2) includes a good introduction and conclusion, (3) uses information from the article stated in your own words plus your own ideas, and (4) follows the rules of writing." The instructions were professionally judged by a panel of writing researchers as appropriately eliciting an informational written response. A previous study indicated that <2% of the variance in students' performance on these prompts was due to differences across prompts (Truckenmiller et al., 2020).

Measures

Students took three different prompts (each using a different source text passage) administered through the WA on 3 separate days within a 2-week period of time. During that same time period, they were administered five subtests of the Test of Written Language, Fourth Edition (TOWL-4; Hammill & Larsen, 2009). At the end of the school year, the school administered the state achievement test in English Language Arts (MSTEP).

Writing Architect

Quantity was the only metric that was automatically scored by the computer. Total words written (TWW) was calculated as the total number of words separated by a space regardless of spelling or order. The remaining measures were human scored within the WA's backend system as detailed below.

Word-Level Errors and Sentence-Level Errors. In the WA's backend scoring system, a student's type-written response to each passage appeared on a screen with different buttons that allowed human scorers to mark errors. Trained research assistants marked word-level errors by using the spelling button to mark any spelling errors and the capitalization button for missing capitalization at the beginning of a sentence, proper names, and the word ‘I'. A copy of the scoring manual with more specific rules and examples is available at https://osf.io/tfvx2.

The computer program totaled the number of each type of error. For word-level errors, the computer divided the number of spelling and capitalization errors by the total number of words for a percent error (similar to percent incorrectly spelled words).

For sentence-level errors, research assistants used (a) the punctuation button for missing punctuation at the end of a sentence (including run-on sentences), missing commas in a series, or a missing comma after an introductory clause; and (b) the syntax button when a word was missing, an unnecessary word was added, there was noun-modifier disagreement, and there were verb-tense errors. Research assistants also used a correct word sequences button to indicate when words were sequenced in syntactically correct order and correct punctuation was included. The computer tallied the total number of sentence-level errors and total number of correct sequences. To calculate sentence-level errors, the computer divided the number of errors by the sum of sentence-level errors and correct sequences for a percent error, which is similar to percent incorrect writing sequences without the spelling errors included. Interrater reliability was calculated on 19% of the samples and was calculated in the same manner as the NAEP using a two-way random absolute agreement intraclass correlation (ICC). The resulting ICC for all errors was 0.94 (95% CI [0.92, 0.96]).

Text-Level Plan. Students' planning was collected via permanent product. Students were given an unlined sheet of blank paper and instructed to plan how they would respond to the prompt. Paper was provided so that students were not restricted by or cued to use specific tools available through a computer application. The paper plan was scored using a 4-point scale of 0–3 to represent the level of structure a student explicitly used prior to writing their composition (0 = *no planning* [student wrote fewer than 5 words], 1 = *no structure* [student simply began writing their first draft], 2 = *general structure* [student used a list of two or more items that may or may not include bullets], and 3 = *specific structure* [student used an organizational strategy such as columns, a graphic organizer like a Venn diagram, or organizational terms such as "introduction," "conclusion," "main idea," "points," "warrants," "claims"]). Interrater reliability was calculated on 20% of the sample using the exact agreement method at 92%.

Typing Fluency. Participants were instructed to type the paragraph appearing at the top of the screen into a text box at the bottom as quickly and accurately as possible and the web application ended administration at 90 s. The paragraph was an extended version (147 words) of the Monroe and Sherman (1966) handwritten paragraph copying task. Student responses were scored as the number of characters correctly typed in 90 s (Graham et al., 1997). Interrater reliability was calculated on 20% of the samples and was perfect, $r = 1.00$.

Writing Quality. The students' final submitted essay response was hand scored for quality using a researcher-developed rubric (Troia et al., 2020). Research assistants scored each of five dimensions on a scale of 0–5 points, for a total scale ranging from 0 to 25. The five dimensions included (a) orients the reader to the purpose of the text effectively and creatively; (b) groups related ideas to enhance text coherence logically and insightfully; (c) provides a concluding sentence or section that follows smoothly from prior ideas; (d) links ideas using words or phrases precisely and effectively for strong cohesion; (e) develops ideas using facts, examples, experiences, descriptive details, and quotes (from source materials as appropriate) that are relevant and impactful; and (f) uses language and vocabulary that is precise, varied, and apt for the type of text. The original rubric also had a sixth dimension for scoring mechanics (i.e., spelling and grammar). We did not include this dimension because it is overaligned with word-level and sentence-level errors being explored in the present study. We wanted to understand the role of word- and sentence-level errors on a measure of quality that did not factor mechanics into the score. The scale used to score each dimension was: 0 = *no evidence of dimensional quality, severely flawed/incomprehensible*; 1 = *minimal evidence of dimensional quality, substantially flawed/difficult to read*; 2 = *some evidence of dimensional quality, notably flawed but readable*; 3 = *adequate evidence of dimensional quality, a few consistent flaws but readable*; 4 = *strong evidence of dimensional quality, some inconsistent flaws/easy to read*; 5 = *excellent evidence of dimensional quality, virtually no flaws/fully comprehensible*. A previous study demonstrated appropriate interrater reliability with correlations above 0.70 (Troia et al., 2020). Interrater reliability in the present study was calculated on 42% of the samples and was calculated in the same manner as the NAEP using a 2-way random absolute agreement intraclass correlation (ICC). The resulting ICC for the total quality score was 0.82 and internal consistency of the six dimensions with the total rubric score was high, α = 0.94. Exact agreement for interrater reliability of the six dimensions was lower than 70% (see Table 3 for interrater reliability of the six domains). Agreement for most domains was consistent with other instances of highly trained raters (i.e., NAEP), but two domains (i.e., purpose and language) were unacceptably low. Therefore, only the total score should be used.

Table 3
Writing quality score interrater agreement.

Quality dimension	Intraclass correlation	Exact agreement	Agreement within 1 point
Coherence	0.79	58%	96%
Cohesion	0.70	51%	97%
Conclusion	0.62	58%	90%
Language	0.73	45%	93%
Purpose	0.61	26%	85%
Support	0.79	60%	94%

Note. Each dimension is scored on a scale of 0–5 points, therefore, the likelihood of exact agreement by chance is 17%. By comparison, NAEP scoring seeks to keep exact agreement on a 6-point scale at 60%, but has many items with agreement between 55%–59%. Source: NAEP 2011 Writing assessment Technical Documentation.

Test of Written Language, Fourth Edition

Students took five subtests of the TOWL-4: Vocabulary, Spelling, Punctuation, Sentence Combining, and Spontaneous Writing. Each subtest produces a norm-referenced score. Age-based norms were used such that each student's raw score performance was compared with their age group (e.g., age 10 years, 0 months to age 10 years, 11 months). The normative scale has a *M* of 10 and *SD* of 3 for the Vocabulary, Spelling, Punctuation, and Sentence Combining subtests. The normative scale has a *M* of 100 and *SD* of 15 for the Spontaneous Writing Index. The TOWL manual reported adequate reliability and validity for the 4th edition norms (Hammill & Larsen, 2009). Within the present sample, the internal consistency reliability for the Spontaneous Writing Index was α = 0.89.

M-STEP Writing

The 2017 MSTEP series of assessments (Michigan Department of Education, 2017) was derived from the Smarter Balanced Assessment Consortium (SBAC) assessments that are used to evaluate students' achievement of grade level expectations. In 2017, 15 states used the SBAC, in whole or in part. The MSTEP writing scaled score in Grades 5 and 8 was calculated from performance on 10 items that consisted of 5 computer-adaptive items about writing organization and purpose (e.g., “Click on two sentences that are not relevant to the writer's argument”), 4 computer-adaptive items on editing writing conventions (e.g., “Edit the sentences by clicking on the sentence that does not use verb tense correctly”), and one 10-point essay that was either a narrative, informational/explanatory, or opinion/argumentative essay. The 10-point essay was scored by assigning up to 4 points for organization and purpose, up to 4 points for development and elaboration, and up to 2 points for conventions. Human raters hired by the state scored the essay. Interrater reliability ranged between 62%–75% perfect agreement; the percentage agreement within 1 point was above 98%. Marginal reliability was reported only for the total English Language Arts score (i.e., reading, writing, listening, and research combined) and was 0.92 in Grade 5 and 0.89 in Grade 8 (Michigan Department of Education, 2017). The passing scaled score for Grade 5 was 1500 and the passing scaled score for Grade 8 was 1800.

Procedures

The WA was group-administered to students in a computer lab or using netbooks in their English Language Arts classroom. All instructions were provided in the web application and were delivered orally by researchers. Students were provided with headphones to listen to the passage presented and a paper packet. The process of reading the passage, planning the response (approximately 3 min), writing the response (maximum time of 15 min), and completing the paragraph copy typing fluency task, plus transitions at the beginning and end of the session, spanned less than one class period (approximately 40 min). The TOWL-4 subtests were administered by researchers in a group format with paper packets. For each classroom, administration of all sessions was conducted within a 1- month period. All classrooms participated in the winter and spring of 2017. The schools group administered the MSTEP in the spring of 2017.

Design and Data Analysis

A counterbalanced form design was used wherein all participating students responded to each of the grade level prompts, but students were randomly assigned to a different order. Truckenmiller et al. (2020) provided a more detailed description of the form design and equating process. Because three prompts were administered to students in a counterbalanced manner within a short time frame, student-level mean scores were used in the current study for all WA variables (i.e., text-level plan, typing fluency, word-level errors, sentence-level errors, WA quality rating, and TWW).

To examine the diagnostic component scores of the WA, path analyses were fit using M*plus* 8.3 (Muth´en & Muth´en, 2019). To address the research questions of structural and substantive validity, models depicted in Figs. 1 and 2 were fitted to the data where text- level plan, word-level errors, sentence-level errors, and typing fluency have direct and indirect relations to MSTEP via overall writing quality and quantity (TWW). Because of the differences in the assessments (passages in the WA assessment as well as MSTEP) between Grades 5 and 8, we fit a separate path model for each grade level.

Results

Evidence of External Validity

The first aim explored external validity evidence, which requires that the scores from the assessment are correlated with gold- standard assessments of the constructs (Messick, 1998). We evaluated the external validity of the four WA component scores with concurrent administration of a norm-referenced assessment of the same constructs and include the correlations in Table 4 (below diagonal for Grade 5, above diagonal for Grade 8). Except for the sentence-level errors at Grade 8, correlations were near the recommended values of 0.60 for demonstrating the level of construct validity needed for use in classroom instructional decisions (NCII, 2019). The correlation of Grade 5 sentence-level errors with TOWL-4 Punctuation was slightly below the threshold at − 0.56, but still met the definition of external validity evidence because the TOWL-4 Punctuation task only measures punctuation whereas sentence- level errors also include errors in syntax. The overall text quality on the WA prompts correlates with the Spontaneous Writing Quotient of the TOWL at 0.65 in Grade 5 and 0.58 in Grade 8. In Grade 5, the significant correlations of word-level errors with TOWL-4 Vocabulary and sentence-level errors with TOWL-4 Sentence Combining suggests that the word- and sentence-level errors capture significant aspects of Vocabulary and Sentence Combining subtest performance. Except for Grade 8 sentence-level errors, we propose that the WA word- and sentence-level errors represent both word- and sentence-level *conventions* (i.e., TOWL-4 Spelling and Punctuation) as well as word- and sentence-level *meaning* (i.e.,TOWL-4 Vocabulary and Sentence Combining) because the WA word- and sentence-level errors represent accuracy, which is a critical aspect to establishing meaning.

The low correlations for sentence-level errors in Grade 8 is likely due to the floor effects observed. Approximately 75% of this sample made <2% errors, indicating that sentence-level errors may not be useful for high-performing Grade 8 classrooms like the ones in our convenience sample.

Table 4
External validity correlations and descriptive statistics.

										Grade 8 Descriptives		
Measure	1	2	3	4	5	6	7	8	9	n	Mean	SD
1. WA Word-level errors		0.19*	− 0.37*	− 0.44*	− 0.65*	− 0.24*	− 0.50*	− 0.32*	− 0.59*	110	0.05	0.04
2. WA Sentence-level errors	0.33*		− 0.16	− 0.09	− 0.30*	− 0.22	− 0.39*	− 0.12	− 0.29*	110	0.02	0.01
3. WA Text-level plan	− 0.15	− 0.16		0.49*	0.44*	0.20	0.37*	0.25	0.35*	110	1.08	0.79
4. WA Quality	− 0.40*	− 0.34*	0.38*		0.45*	0.42*	0.46*	0.31*	0.58*	110	13.71	4.33
5. TOWL Spelling	− 0.60*	− 0.47*	0.26*	0.45*		0.46*	0.66*	0.44*	0.58*	102	12.18[a]	2.58
6. TOWL Vocabulary	− 0.38*	− 0.34*	0.21*	0.58*	0.56*		0.37*	0.34*	0.40*	105	12.14[a]	2.63
7. TOWL Punctuation	− 0.52*	− 0.56*	0.28*	0.46*	0.70*	0.54*		0.23	0.53*	102	12.29[a]	2.63
8. TOWL Sentence Combining	− 0.39*	− 0.46*	0.26*	0.46*	0.44*	0.43*	0.54*		0.45*	102	13.09[a]	2.61
9. TOWL Spontaneous index	− 0.56*	− 0.36*	0.32*	0.65*	0.59*	0.44*	0.61*	0.46*		102		16.66
Grade 5 Descriptives												
n	175	175	175	175	167	168	167	164	167			
Mean	0.07	0.04	1.31	11.09	10.98[a]	8.44[a]	11.09[a]	11.58[a]	113.56[b]			
SD	0.06	0.03	0.88	4.43	2.29	2.91	2.24	4.29	16.9			

Note. Grade 5 correlations are below the diagonal and Grade 8 correlations are above the diagonal. All correlations are p < .01 unless noted. NS = not significant, TOWL = Test of Written Language, 4th Edition; WA = Writing Architect. [a] For these TOWL subtests, the age-based normative scaled score mean is 10 and the standard deviation is 3. [b] For the TOWL Spontaneous Index, the age-based normative index mean is 100 and the standard deviation is 15. $^*p$ < .01, two-tailed.

Preliminary Analyses and Descriptive Statistics for Structural and Substantive Validity

Structural and substantive validity were examined within one path model. Preliminary analyses were conducted separately for Grades 5 and 8 for all the variables included in the model. For both grade levels, missingness was minimal (1.42% for Grade 5 [11 observations] and <0.16% for Grade 8 [two observations]). For Grade 5, Little's test revealed data were missing completely at random, χ^2 (12) = 9.82, p = .63. Thus, we used Full Information Maximum Likelihood (FIML) method to address the missingness. For Grade 8, Little's test indicated data were not missing completely at random, χ^2 (6) = 16.025, p = .01. Test of analyses of variance suggested that two missing cases performed significantly lower on MSTEP and other measures than students with complete data. To address missingness, we performed multiple imputation (creating 10 imputed datasets) and the results were identical to those of FIML findings. Therefore, we report results from the model using FIML.

In both grade levels, 4–5 univariate outliers were detected using Tukey's (1977) method (i.e., three inter-quartile below the 0.25 percentile or above the 0.75 percentile) and they were winsorized to the outer fence values (Reifman & Keyton, 2010). Once outliers were winsorized, we did not identify any multivariate outliers using the blocked adaptive computationally efficient outlier nominators (BACON) algorithm in Stata 13.0 (StataCorp, 2013) proposed by Billor et al. (2000). After winsorizing the outliers, all the variables demonstrated appropriate univariate distributional characteristics (see Table 5) as indicated by skewness (±2) and kurtosis values (< 7; West et al., 1995). However, both Grade 5 and 8 data deviated from multivariate normality (Mardia mSkewness = 14.19, χ^2 (165) = 337.28, $p < .05$, Mardia mKurtosis =106.65, χ^2 (1) = 10.26, $p < .05$, for Grade 5; Mardia mSkewness = 17.03, χ^2 (84) = 317.16, $p < .01$, Mardia mKurtosis =75.47, χ^2 (1) = 33.34, $p < .01$, for Grade 8). Thus, analyses were conducted using robust variance in maximum likelihood estimation using MLR in M*plus* given the nonnormality of the data.

Additionally, clustering of students at the teacher-level was considered. In Grade 5, there were five teachers, with teacher-level intraclass correlations (ICCs) ranging from 2% to 10%. Due to having a small number of clusters (i.e., five teachers) relative to the number of parameters estimated in the model, standard errors of estimates were unreliable. Thus, we were not able to account for teacher-level variances in a multi-level analysis. Instead, we added dummy-coded teacher variables as fixed effects to the model.

In Grade 8, there was only one teacher who taught five sections of students. The ICC for MSTEP at the section-level was <1% (ICC = 0.002). Therefore, to be parsimonious and consistent with the Grade 5 model, section assignment was not included as fixed effects in the model.

Table 5 shows descriptive statistics, including the *M*, *SD*, minimum, maximum, skewness, and kurtosis for each variable, as well as bivariate correlations between the variables. As expected, word-level and sentence-level errors had negative, moderate correlations with writing outcome measures in both grades (− 0.31 to − 0.51 for Grade 5 and − 0.05 to − 0.45 for Grade 8), whereas text-level plan and typing fluency were positively correlated with writing outcome measures in both grades (0.27–0.60 for Grade 5 and 0.31–0.49 for Grade 8). Quantity (i.e., TWW) and rating of writing quality were highly correlated with each other (above 0.90 in both grades) and they were moderately correlated with MSTEP performance (above 0.51 for Grade 5 and 0.63 for Grade 8).

Table 5

Descriptive statistics and correlations for variables in the path models.

								Grade 8 Descriptives					
Measure	1	2	3	4	5	6	7	Mean	SD	Skew	Kurtosis	Min	Max
1. Word-error		0.20*	− 0.37*	− 0.47*	− 0.33*	− 0.44*	− 0.45*	0.05	0.04	1.84	3.06	0.00	0.18
2. Sentence-error	0.33*		− 0.16	0.08	− 0.05	− 0.09	− 0.18	0.02	0.00	0.85	0.39	0.00	0.06
3. Text-level plan	− 0.15	− 0.16*		0.22*	0.48*	0.49*	0.46*	1.08	0.78	0.18	− 0.78	0.00	3.00
4. Typing Fluency	− 0.45*	− 0.22*	0.16*		0.36*	0.36*	0.31*	189.25	65.40	1.21	2.55	69	426
5. Quantity	− 0.31*	− 0.32*	0.32*	0.60*		0.92*	0.63*	165.42	70.30	0.49	0.41	22	412
6. Quality	− 0.40*	− 0.34*	0.38*	0.56*	0.90*		0.68*	13.71	4.31	−0.16	0.18	2.00	24.67
7. MSTEP	− 0.51*	− 0.46*	0.27*	0.42*	0.51*	0.56*		1808.87[b]	27.75	−0.46	−0.17	1736	1857
Grade 5 Descriptives													
Mean	0.07	0.04	1.31	114.77	119.89	11.09	1512.93[a]						
SD	0.05	0.03	0.88	43.97	57.21	4.41	27.44						
Skew	1.52	1.26	0.44	0.95	0.64	− 0.20	− 0.54						
Kurtosis	2.43	2.04	− 0.62	2.03	0.64	− 0.45	0.07						
Min	0.00	0.00	0.00	37.67	11.00	0.00	1414						
Max	0.28	0.19	3.00	322.00	337.67	21.00	1560						

Note. Grade 5 correlations are below the diagonal and Grade 8 correlations are above. All correlations are significant at p < .05 unless otherwise noted. NS = not significant. MSTEP = Michigan Student Test of Educational Progress.

[a] The passing scaled score for Grade 5 was 1500.

[b] The passing scaled score for Grade 8 was 1800.

[*] p < .05, two-tailed.

Model for Evidence of Structural and Substantive Validity

Path analyses were conducted to obtain evidence of structural validity (i.e., the four component scores were related to writing quality and quantity) and substantive validity (i.e., scores on the WA generalize to performance on the state test). Prior to fitting the path models, given the large scaling differences across the variables as shown in Table 5, variables were linear transformed prior to the analyses (e.g., word- and sentence-level errors were multiplied by 100). The fitted models were just-identified models; thus, model fit statistics were not evaluated. Direct effects, indirect effects, and total effects from the models are summarized in Table 6. Direct effects demonstrate how students' skills measured on the WA are directly tapped by the MSTEP. Indirect effects are the sum of the transmission of writing skills through writing quality and writing quantity to performance on the MSTEP.

In Grade 5, the quality of WA prompt responses was predicted by text-level plan ($\beta = 0.26, p <. 01$), sentence-level errors ($\beta = - 0.15, p = .035$), and typing fluency ($\beta = 0.43, p < .01$) after controlling for teacher effects. Similarly, the quantity (i.e., TWW) also was predicted by text-level plan ($\beta = 0.21, p <. 01$), sentence-level errors ($\beta = - 0.15, p = .025$), and typing fluency ($\beta = 0.51, p < .01$). When predicting the MSTEP outcome, the only significant predictors that had direct effects were word- and sentence-level errors ($\beta = - 0.24, p <.01$ and $\beta = -0.24, p <.01$, respectively). The negative correlations for word- and sentence-level errors were expected given that students with fewer errors performed higher on more general writing metrics.

Table 6

Direct, indirect, and total effects of writing architect components through writing quality and quantity on M-STEP.

Variable	Direct	se	*p*-value	Indirect	se	*p*-value	Total	se	*p*-value
	Grade 5								
World-level Error	**-0.24***	0.07	< 0.001	− 0.02	0.03	0.610	**-0.26***	0.07	< 0.001
Quality				− 0.03	0.02	0.345			
Quantity				0.00	0.01	0.576			
Sentence-level Error	**-0.24***	0.07	0.001	**-0.05***	0.03	0.030	**-0.29***	0.08	< 0.001
Quality				− 0.04	0.03	0.241			
Quantity				− 0.01	0.02	0.394			
Text-level plan	0.09	0.06	0.176	**0.08***	0.03	0.011	**0.17***	0.06	0.010
Quality				0.07	0.04	0.245			
Quantity				0.02	0.03	0.351			
Typing Fluency	0.05	0.08	0.568	**0.15***	0.04	0.000	**0.20***	0.07	0.010
Quality				0.11	0.06	0.227			
Quantity				0.04	0.06	0.342			
	Grade 8								
World-level Error	− 0.14	0.09	0.131	− 0.09	0.06	0.125	**-0.23***	0.09	0.011
Quality				− 0.08	0.05	0.139			
Quantity				− 0.01	0.01	0.648			
Sentence-level Error	− 0.11	0.07	0.119	− 0.03	0.05	0.584	− 0.14	0.09	0.111
Quality				− 0.02	0.04	0.549			
Quantity				0.00	0.01	0.788			
Text-level Plan	0.13	0.08	0.121	**0.19***	0.04	0.000	**0.32***	0.08	< 0.001
Quality				**0.15***	0.07	0.039			
Quantity				0.04	0.06	0.475			
Typing Fluency	0.03	0.08	0.703	0.10	0.05	0.067	0.13	0.09	0.149
Quality				0.07	0.06	0.184			
Quantity				0.03	0.04	0.499			

Note. se = standard error. $^{*}p$ < .05.

The estimated relations between component skills, proximal writing performance (quality and quantity), and distal writing performance (MSTEP score) are illustrated in Fig. 1 for Grade 5. Although word-level errors had only a direct effect on MSTEP performance, sentence-level errors had a statistically significant total indirect effect ($\beta = -0.05$, $p = .044$). Additionally, although text-level plan did not show a direct relation to MSTEP performance, the total indirect effect of text-level plan was statistically significant ($\beta = 0.08$, $p < .01$). The same pattern was true for typing fluency; the total indirect effect of typing fluency was statistically significant ($\beta = 0.15$, $p < .01$). Total effects, including both direct and indirect effects of word- and sentence-level errors, on the MSTEP were substantial (− 0.30 for both types of errors) and larger than text-level plan (0.16) or typing fluency (0.19).

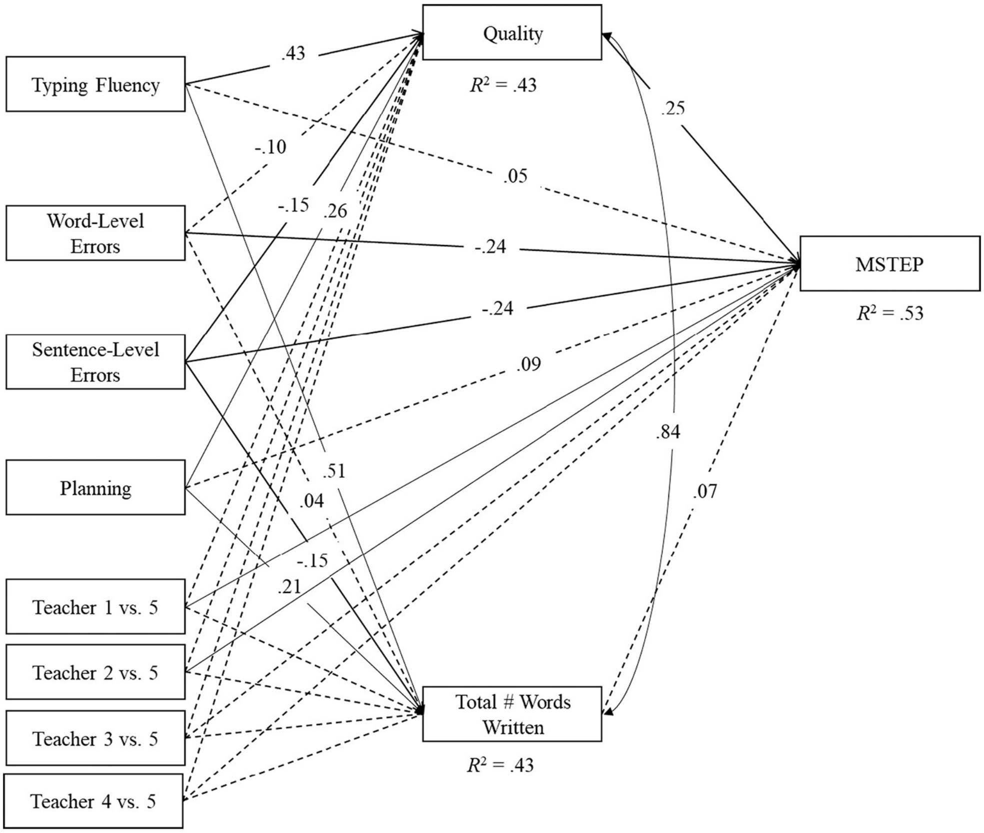

Fig. 1. Grade 5 Path model of evidence for structural and substantive validity.
Note. Paths with bolded lines are statistically significant, $p < .025$. Non-significant paths are represented by dotted lines. MSTEP = Michigan Student Test of Educational Progress.

Evidence of Generalizability Validity

The Grade 8 model is illustrated in Fig. 2. The quality of the Grade 8 WA prompt responses was predicted by text-level plan ($\beta = 0.36, p <. 01$) and word-level errors ($\beta = - 0.20, p = .02$), whereas the quantity (total number of words written) was predicted by text- level plan ($\beta = 0.39, p <. 01$) and typing fluency ($\beta = 0.25, p = .02$). When predicting the MSTEP outcome, the only significant predictor that had a direct effect was the quality rating ($\beta = 0.41, p =. 04$). Text-level plan also showed an indirect relation to MSTEP with a total indirect effect of 0.19 ($p < .01$), specifically through quality (0.15, $p = .04$). However, neither typing fluency nor word- or sentence-level errors had indirect effects on MSTEP performance (0.10, − 0.09, and − 0.03, respectively). Total effects, including both direct and indirect effects of text-level plan on MSTEP, were substantial (0.32) and larger than other indicators.

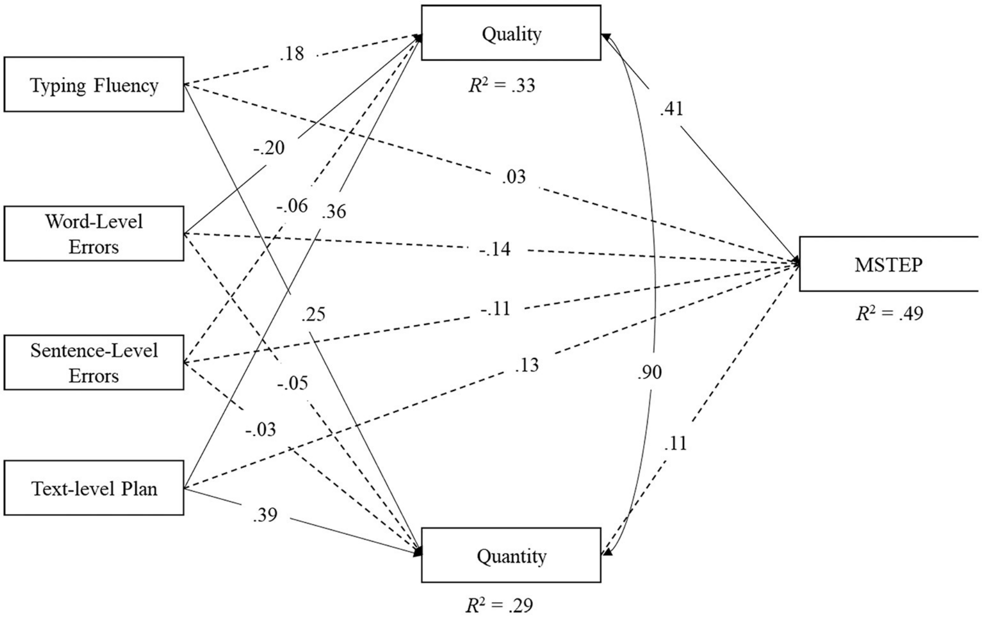

Fig. 2. Grade 8 Path model of evidence for structural and substantive validity.
Note. Paths with bolded lines are statistically significant, $p < .05$. Non-significant paths are represented by dotted lines. MSTEP = Michigan Student Test of Educational Progress.

Discussion

In classrooms, assessment derives its utility by informing the educator's next instructional steps. Teachers' use of writing quality rubrics has guided effective instructional practices like individualized detailed feedback and teaching self-regulated writing strategies (Graham et al., 2016; Rouse & Kiuhara, 2017). However, students' performance on writing rubrics can be constrained by their development in transcription and accurate word-, sentence-, and discourse-level writing skills that overwhelm limited working memory capacity (Berninger et al., 2002; Kellogg et al., 2013; McCutchen, 2011; Troia et al., 2019). Interventions for transcription skills and writing accuracy positively impact writing quality (Graham et al., 2016; Graham & Santangelo, 2014) and teachers seek assessment to determine which students need this kind of supplemental instruction (Applebee &

Langer, 2011; Graham et al., 2014). In this study, we gathered several sources of evidence to expand the utility of a classroom assessment to meet instructional decision- making needs.

What Are We Measuring in the Classroom?

External Validity

Obtaining a valid measure of underlying skills with a classroom assessment is a significant challenge. Our study provides several sources of evidence that four diagnostic component skills (i.e., typing fluency, word-level errors, sentence-level errors, and text-level plan) can be measured in one task, especially for students in Grade 5. Evidence is provided that word accuracy and sentence accuracy represent spelling and punctuation abilities for students in Grade 5 and spelling in Grade 8 (i.e., evidence of external validity). The significant direct paths of word- and sentence-level errors to the Grade 5 MSTEP performance further suggests that word- and sentence- level errors represent students' spelling and grammar skills that are tapped by both a distal outcome (i.e., MSTEP) and a proximal classroom task (i.e., writing quality).

The lack of generalizability for sentence-level errors in Grade 8 contributes to mixed evidence about later sentence development and its impact on writing. Our Grade 8 sample was in the high average range compared to national norms in writing (mean TOWL-4 Spontaneous Writing Index = 110) and they made very few sentence-level errors (*M* = 2%, *SD* = 1%). Another study in Grades 7 and 8 found that sentence-level accuracy was significantly but weakly correlated with writing quality (*r* = 0.17; Limpo et al., 2017). Limpo et al. also combined their accuracy measure with a sentence combining task and found that this composite was a small but significant factor for writing quality. In our sample, the students' Sentence Combining scores as measured by the TOWL-4 were significantly related to writing quality. It is likely that accuracy of sentence writing is not the best measure to represent older students' variability in their sentence-level skills. Rather, complexity metrics may be more appropriate targets for assessment and instruction in the later grades (e.g., see Troia et al., 2019; Truckenmiller & Petscher, 2020). To meet the assessment efficiency goal of measuring sentence-level characteristics within the same task, a robust complexity metric needs to be identified. Some researchers have suggested that indices generated in Coh-Metrix (McNamara et al., 2005) show promise for measuring sentence-level complexity, but more research is needed to find out what specific writing skills (and corresponding areas to target for instruction) those Coh-Metrix indices actually represent for students still developing their writing skills (Troia et al., 2019; Wilson et al., 2017).

Structural Validity

In Grade 5, the four component skills explained 43% of the differences in students' writing quality and amount of text produced on the WA task. This suggests that when educators see the large variability that occurs in their students' written compositions (Graham et al., 2014), almost half of it is due to strengths or weaknesses in typing fluency, word- and sentence-level errors, and text-level plan. This result is very similar to Limpo et al. (2017) who evaluated these four component skills using separate measures and found that they accounted for 46% of the variance in writing quality, even though the study included slightly older students and assessed writing in the opinion genre. Clearly, these four component skills are not the only ones that influence writing performance, but this evidence suggests that they cannot be ignored when educators are making plans to improve students' writing through instruction.

In Grade 8, the four component skills represent 33% of the variance in students' writing quality, which is less than what was found in a previous study (Limpo et al., 2017) but still indicates substantive information to be considered for eighth graders. The Grade 8 model demonstrated patterns that are more consistent with what one would anticipate for students who had mastered foundational writing skills and had freed cognitive capacity to focus more on the organizational and meaning aspects of what they were writing (e.g., Kellogg et al., 2013; Olive, 2014; O'Rourke et al., 2018). At this point in student writing development, additional variables should be included in assessment to help determine instructional needs in sentence coherence and variability of sentence type (which are measured in the quality rubric). These variables might include other factors at the discourse level as well as complexity variables at the word- and sentence-levels of language (Koutsoftas & Petersen, 2017; McNamara et al., 2009; Silverman et al., 2015; Troia et al., 2020). Future studies using additional complexity variables could increase the structural validity of the WA for more advanced students.

Because the purpose of this study was to find validity evidence for specific types of assessment data that each have separate instructional implications, we did not evaluate the dimensionality of relations between the four diagnostic component skills. Empirical models are beginning to explore the interrelated roles of such variables in the context of other cognitive components of writing at different ages. For example, in first grade, spelling, handwriting, and discourse structure (comprising vocabulary and grammar) were the primary predictors of written composition performance (Kim & Schatschneider, 2017). These three constructs were found to be separable but correlated factors. In a middle school example, the best-fitting model included handwriting fluency mediated by planning and spelling mediated by translating (i.e., sentence combining and syntactic correctness) to predict written composition performance (Limpo et al., 2017). Although these structural relationships will be important for understanding student development of writing ability, further study will be needed to establish how they can be interpreted in a practical way to guide classroom instruction (McCardle et al., 2018).

Substantive Validity

Not only do the four diagnostic components influence the quality and length of the written composition, but these components also generalize to a more distal outcome of writing achievement. The model in Fig. 1 provides evidence that, for Grade 5, the influence of students' typing fluency, text-level plan, and errors with sentence structure are transmitted through their effects on informational composition quality and quantity to predict distal writing achievement. Grade 5 students' errors with words and sentences also directly predicted distal writing achievement. Our models predicted approximately half of the variance in end-of-year writing achievement on the MSTEP, indicating that the informational writing task and derived metrics we chose have significant relevance to student outcomes, even when those outcomes may vary in genre. In Grade 8, writing quality directly predicted distal writing achievement and both text-level plan and spelling had significant total effects on MSTEP performance when transmitted through writing quality. Although the model still predicted almost half of the variance in writing achievement at Grade 8, typing fluency, sentence-level errors, and quantity on the informational writing task were no longer the driving factors of distal writing achievement. This finding is consistent with most models in middle school that find word- and text-level indicators to be the most robust (e.g., Abbott et al., 2010).

Future Evidence for Generalizability

The present study provides evidence for two groups consisting of a demographically relatively homogenous sample of Grade 5 students and a high average achieving sample of Grade 8 students. Although these samples varied as much as a nationally normed sample (i.e., the *SD*s on the TOWL-4 for the sample were similar to national norms), further study with different populations is needed. Specifically, the relative influence of different components may be different for higher- and lower-performing students (Berninger et al., 1994; Kent & Wanzek, 2016; McCardle et al., 2018; Truckenmiller et al., 2021).

Implications for Practical Use and Future Research

The evidence provided so far shows promise for content, external, substantive, structural, and generalizability validity of the WA assessment and therefore shows promise for next steps in exploring consequential validity. Although these types of validity are necessary, consequential validity is most needed by schools (Gersten et al., 1995; Messick, 1998). Gersten et al. (1995) illustrated rather convincingly that typical construct and predictive validity evidence are not enough to promote effective *use* of assessment in schools by educators to promote academic achievement. They illustrate this point with the history of reading assessment. In the late 20th century, researchers found that, although assessment of discrete auditory and visual processing skills predicted reading achievement, the assessment of these skills was not useful because evidence did not support differentiating auditory and visual instruction as a means of improving reading outcomes. Therefore, assessment must identify components that educators can directly translate to instruction. Studies should be conducted to determine what educators do with the assessment information, including both the intended uses and especially the unintended uses. Gersten et al. (2020) used a mixed methods study to evaluate how educators use diagnostic assessment of malleable components of mathematics. In the spirit of finding malleable component skills of writing, we next detail the intended use of the WA diagnostic scores and recommend future consequential validity research that collects evidence of how educators would use the information as well as identifying unintended uses.

Typing Fluency

The consequential validity of typing fluency scores is relatively transparent: to provide effective typing instruction for those who need it so that they can write without devoting significant cognitive resources to typing. Handwriting interventions have been effective for middle school students (Santangelo & Graham, 2016) and there is reason to believe that similar interventions for building typing fluency may follow a similar pattern. In fact, because there are more students still building fluency with typing than building fluency with handwriting in middle school, typing instruction may be more widely needed (Berninger et al., 2009; Feng et al., 2019).

Because quantity of writing is so interconnected with quality of writing in early middle school ($r = 0.90$), difficulties with typing fluency cannot be ignored as simply a nuisance factor. In fifth grade, typing fluency demonstrated a stronger relationship with writing quality than text-level plan, indicating that foundational skills continue to play an outsized role in composition, even though most instruction has moved to the text level at this point (Graham et al., 2014). This finding is consistent with other studies (e.g., Troia et al., 2020). Conversely, by the end of middle school, typing fluency only predicted writing quantity. Although writing quantity and quality are still interconnected ($r = 0.92$) for eighth graders, typing fluency does not have a significant direct, indirect, or total effect on writing achievement. As a point of reference, Grade 8 students typed >1.5 times faster on average than Grade 5 students on average. It is possible that typing fluency at Grade 8 reaches a mastery threshold for most students, although a mastery threshold is difficult to determine because published norms for typing fluency are not available—a mastery criterion is needed for future research to promote consequential validity.

Word-Level Errors

In our Grade 5 sample, word-levels errors were more strongly correlated with the MSTEP than with the WA quality performance. This resulted in our model showing significant direct effects of word-level errors to the MSTEP, but not significant indirect effects (through WA quality). It is possible that scoring of the MSTEP is overly influenced by presentation effects, wherein scorers have difficulty assigning higher scores for the students' ideas when the legibility and spelling is poor (Graham et al., 2011). It is also possible that the MSTEP multiple-choice items measured constructs impacted by spelling. Regardless, our study provides significant content and generalizability evidence for spelling. In Grade 8, there was a significant total effect of word-level errors on distal writing achievement, indicating that word-level accuracy is important via multiple pathways throughout middle school. These findings are consistent with other research that indicates spelling plays an important but changing role throughout writing development (Abbott et al., 2010; Graham et al., 1997; Kim et al., 2015; Olinghouse, 2008).

There is ample evidence to support the external validity of using assessment of word-level errors to recommend spelling instruction and that it has impacts on proximal and distal writing outcomes, especially in Grades 1–6 (Graham et al., 2016; Graham & Santangelo, 2014). For later grade levels, researchers have suggested integrating spelling instruction with other writing instruction to have larger effects on distal writing achievement (e.g., Graham & Santangelo, 2014). Evidence from the current study aligns well with that recommendation. Our models show direct effects of spelling in fifth grade and indirect effects in eighth grade.

Furthermore, we chose to label the spelling errors as word-level errors instead of just spelling as an attempt to increase the consequential validity of signaling other types of instruction. Spelling instruction, when integrated with vocabulary and morphology instruction, has been effective not just for middle school students who are struggling, but also for all students as word lengths increase and greater diversity of discipline-specific vocabulary appears in texts (Goodwin & Ahn, 2010; Wright & Cervetti, 2017). The word-level errors in our study had a small but significant relationship with vocabulary in Grade 5 and a significant but negligible relationship in Grade 8. This suggests that word-level errors could signal other dimensions of word-level knowledge beyond spelling (e.g., morphological knowledge, facility with academic vocabulary) for middle grade students, but other measures of word complexity (beyond accuracy) may better differentiate written composition performance (Troia et al., 2019).

A future research study might include a measure of complexity to determine if the complexity dimension further differentiates student instructional needs. Consequential validity can be explored by examining whether accuracy and complexity scores help educators focus their instruction on word accuracy and complexity.

Sentence-Level Errors

As anticipated, sentence-level errors played a significant unique role in informational writing quantity and quality in Grade 5 and had a direct effect on end-of-year writing achievement. It could be argued that the direct effect of sentence-level errors is due to heavier weight placed on the multiple-choice sentence editing items on the MSTEP and that sentence-level errors would not have an impact on written composition outcomes that did not have these types of items. However, the Grade 8 MSTEP also had these items, and sentence- level errors did not have an impact at that grade, likely due to very few sentence-level errors committed by our sample of eighth-grade students. Other studies have suggested that there is a small but significant relationship between sentence-level accuracy and written composition outcomes (e.g., Limpo et al., 2017; Truckenmiller & Petscher, 2020).

Given the significant role of sentence-level accuracy in early middle school writing development, effective instruction in this area is warranted. Intuitively, educators may look to grammar instruction to bolster sentence-level accuracy. However, grammar instruction with parts of speech or diagramming sentences has not been demonstrated to be effective (e.g., Graham et al., 2012). Rather, proofreading strategy instruction, editing checklists, and helping students to understand the impact of these errors on readers' comprehension can improve sentence-level accuracy (McNaughton et al., 1997). Future studies of educators' use and misunderstandings of a sentence-level accuracy metric will be needed to establish the parameters for consequential validity (Gersten et al., 2020).

Instruction on sentence construction and sentence combining have been effective for improving sentence writing outcomes in Grades 4–11 (e.g., Datchuk & Kubina, 2012). In our Grade 5 sample, the moderate correlation between sentence-level errors and students' Sentence Combining scores suggests that our assessment tool could be used to recommend sentence combining instruction, but future studies that connect the assessment with intervention in sentence combining are needed for evidence of external validity.

By eighth grade, sentence-level errors were not a significant unique predictor of writing quantity or quality. Our Grade 8 sample demonstrated mastery of sentence accuracy because they had very few errors and virtually no variability in the number of errors. Confirmation of the Grade 8 sample's mastery of sentence-level skills also is confirmed by the results of the nationally normed TOWL-4 results. On average, our sample performed more than one standard deviation above the national mean in Sentence Combining (see Table 4), suggesting that a different metric may be needed for more advanced writers. Revision may be something to consider. Revision instruction that incorporates sentence- and discourse-level revisions has been effective for Grade 8 students who struggle in writing (De La Paz et al., 1998). However, revision has been difficult to operationalize in assessment and needs further exploration in substantive validity studies.

Text-Level Plan

Discourse-level performance is often represented by measures (e.g., total number of words written or ideas produced) that are highly dependent on many skills. In our study, we tried to isolate the discourse level with the text-level plan score. Correlations confirmed that the text-level plan score was not related to word- or sentence-level errors (see Table 4) and had low but significant positive correlations with other discourse-level measures in the study (i.e., WA quality, WA quantity, TOWL Spontaneous Writing Index, and MSTEP). Furthermore, when considering the other three diagnostic scores in the WA assessment, text-level plan had a unique contribution to both the quality and quantity of written text and had the highest contribution of the four components at Grade 8. Interestingly, text-level plan did not have a direct path to the MSTEP writing score, suggesting that text-level plan may be situationally dependent. In other words, when the student plans, it has an impact, but the student may choose whether to plan at each assessment. Taken together, these results indicate that planning sophistication does isolate some aspect of discourse-level writing that is important.

A text-level plan score of 2 indicates that the students had some thoughts about the points that needed to be included in the text, whereas a score of 3 indicates that students had a clear conceptualization of the structure of the text they wanted to write, and a score of 1 reflects simple drafting of some text that was copied over for a composition. In Grade 5, 62% of our high-performing sample had scores of 2 or 3 and in Grade 8, 68% had scores of 2 or 3. This suggests that at least a third of the middle school students in this study engaged in "knowledge telling" (Bereiter & Scardamalia, 1987) and did not have an explicit a priori plan for the structure of their texts. Furthermore, the average performance on this measure did not change much from fifth to eighth grade (although longitudinal research is needed to confirm this).

A clear implication of the WA text-level plan score is to teach students to determine the type of informational text they need to write (e.g., compare/contrast, description, problem/solution) and use the structural components to generate the relevant ideas they need to include (e.g., De La Paz, 2007; Hebert et al., 2018). Prewriting activities with graphic organizers and text structure instruction are evidence-based practices with moderate effect sizes on writing outcomes (Graham et al., 2012). Future intervention research is needed to determine if educators share this interpretation and if the text structure instruction for students with scores of 0 and 1 result in improved writing quality (i.e., evidence of external and consequential validity).

Limitations

There are at least three sources of measurement error in the WA that we attempted to address. Interrater reliability for writing quality is typically lower than the reliability that education researchers expect with other academic skills (Calfee

& Miller, 2007) and the reliability obtained in this study is no exception. However, writing quality scores are consistently used to represent the construct of writing more than other measures of written composition (Kent & Wanzek, 2016); therefore, quality continues to be the standard for evaluating written composition, however imperfect. For example, the NAEP is considered a gold standard for measuring writing outcomes and even after extensive training, they find that exact agreement on individual items of a writing rubric fall below 60%. In the present study, we attempted to mitigate lower reliability by using only the total score and using the students' mean performance on three different prompts. The mean was used to address reliability of quality scoring as well as the unknown situational variables that contribute to error for each data point.

One obvious contributor to error across different informational passage prompts is the interaction of student's reading decoding and comprehension abilities with different prompts. A previous study (Truckenmiller et al., 2020) evaluated the amount of variance due to the specific prompts in the WA using a counterbalanced design. The variance due to the passage text (controlling for individual student differences) in the current passage set was 2% of the total variance in writing performance. This suggests that the impact of the different prompts was minimal. We also attempted to minimize the impact of variation in student's decoding skills by having the passages read aloud by a human voice. However, we cannot estimate the extent to which decoding affected students' writing performance.

Variation of each student's language comprehension abilities likely impacted their written composition performance and it is not possible to estimate the extent of this impact. A future study that measures a verbal summary (e.g., oral retell) would be able to identify the extent to which comprehension is transferred to writing. However, we do not view variation in comprehension as completely construct-irrelevant to writing performance. If writing is the product of transcription and idea generation and language formulation, we should expect to see the generated ideas from the passage in the written composition. Studies show that a majority of the variance in written composition can be predicted by reading comprehension (e.g., Truckenmiller & Petscher, 2020). The consensus is that written composition and reading comprehension are conceptually, theoretically, and functionally tapping related constructs and this understanding has important implications for instruction (Berninger et al., 1994; Graham & Harris, 2017; Kent & Wanzek, 2016; Kim, 2020; Kim & Schatschneider, 2017). For example, Graham and Harris, in an aggregation of meta-analyses, found that instruction in written composition has one of the largest effect sizes for facilitating reading comprehension ability.

Finally, student performance in writing cannot be divorced from the instructional environment and sociocultural environment (Graham, 2018) in which students are writing. The instructional environment was not measured in the present study. We were also limited in our ability to account for clustering at the classroom level in Grade 5. We were able to account for only the teacher effects but not at the classroom level (peer effects) which possibly inflated Type I error rates (e.g., McCoach & Adelson, 2010). Yet, we note that classroom effects will likely be minimal and smaller than the teacher effects based on a prior study (Burke & Sass, 2013) and our data. As was the case in our Grade 8 sample, Burke and Sass (2013) estimated the peer effects for reading and mathematics to be very small in Grades 4 to 5 (0.03).

Conclusion

Most assessments used in schools have evidence of validity for narrow purposes (e.g., identifying who needs instruction). However, instructional decisions in schools are complex and this complexity needs to be considered when assessment validity evidence is gathered. Educators need actionable assessment information that directly connects malleable skills with evidence-based instruction (Gersten et al., 2020; McKeown et al., 2019). This study illustrates a task, much like typical classroom writing tasks, that provides different assessment data for different purposes and the type of research evidence needed to support diagnostic decisions. The four diagnostic component skills evaluated in this study made up approximately 40% of how fifth grade students vary in their writing quality and 33% of how eighth grade students vary, which then explains approximately 50% of the variation in performance on an end- of-year writing achievement test for both grades. This has clear implications for instructional targets throughout middle school. Explicit instruction and guided practice with specific skills like spelling, typing, and sentence combining to automatize these processes reduce demands within working memory. Once these processes are automatized, they are less likely to constrain written composition and students can devote cognitive resources to writing to learn content area concepts and information. Although foundational writing skills are not typically evident as a target for instruction in middle school, the use of the informational writing prompts in our study (a perhaps more difficult writing task than the more commonly studied narrative genre) may have drawn out students' dysfluency with producing the more difficult academic words required in informational writing. Our primary conclusion is that this assessment tool is a starting point for a classroom assessment to measure and inform supplemental instruction on some of the components of writing that students certainly need to be successful in middle school (Grade 5) and prepared for high school writing (Grade 8). Further research will be needed to confirm that the use of assessment scores to select specific instruction has an impact on student achievement, as well as identifying other meaning-related variables for assessment, and any unintended decisions that negatively impact diverse groups of students.

References

Abbott, R. D., Berninger, V. W., & Fayol, M. (2010). Longitudinal relationships of levels of language in writing and between writing and reading in grades 1 to 7. *Journal of Educational Psychology, 102*, 281–298. https://doi.org/10.1037/a0019318.

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

Applebee, A. N., & Langer, J. A. (2011). A snapshot of writing instruction in middle schools and high schools. *English Journal, 100*, 14–27.

Bangert-Drowns, R. L., Hurley, M. M., & Wilkinson, B. (2004). The effects of school-based writing-to-learn interventions on academic achievement: A meta-analysis. *Review of Educational Research, 74*(1), 29–58. https://doi.org/10.3102/00346543074001029.

Beers, S. F., & Nagy, W. E. (2011). Writing development in four genres from grades three to seven: Syntactic complexity and genre differentiation. *Reading and Writing, 24*(2), 183–202. https://doi.org/10.1007/s11145-010-9264-9.

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Lawrence Erlbaum.

Berman, R. A., & Verhoeven, L. (2002). Cross-linguistic perspectives on the development of text-production abilities in speech and writing. *Written Language & Literacy, 5*(1), 1–43. https://doi.org/10.1075/wll.5.1.

Berninger, V., & Wolf, B. (2009). *Helping students with dyslexia and dysgraphia make connections: Differentiated instruction lesson plans in reading and writing*. Baltimore, MD: Paul H. Brookes.

Berninger, V. W., Abbott, R. D., Augsburger, A., & Garcia, N. (2009). Comparison of pen and keyboard transcription modes in children with and without learning disabilities. *Learning Disability Quarterly, 32*(3), 123–141. https://doi.org/10.2307/27740364.

Berninger, V. W., Cartwright, A. C., Yates, C. M., Swanson, H. L., & Abbott, R. D. (1994). Developmental skills related to writing and reading acquisition in the intermediate grades. *Reading and Writing, 6*(2), 161–196.

Berninger, V. W., & O'Malley May, M. (2011). Evidence-based diagnosis and treatment for specific learning disabilities involving impairments in written and/or oral language. *Journal of Learning Disabilities, 44*(2), 167–183. https://doi.org/10.1177/0022219410391189.

Berninger, V. W., Vaughan, K., Abbott, R. D., Begay, K., Coleman, K. B., Curtin, G., … Graham, S. (2002). Teaching spelling and composition alone and together: Implications for the simple view of writing. *Journal of Educational Psychology, 94*(2), 291–304. https://doi.org/10.1037/0022-0663.94.2.291.

Berninger, V. W., & Winn, W. D. (2006). Implications of advancements in brain research and technology for writing development, writing instruction, and educational evolution. In C. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 96–114). New York: Guilford.

Billor, N., Hadi, A. S., & Velleman, P. F. (2000). BACON: Blocked adaptive computationally efficient outlier nominators. *Computational Statistics & Data Analysis, 34*(3), 279–298. https://doi.org/10.1016/s0167-9473(99)00101-2.

Bourdin, B., & Fayol, M. (2002). Even in adults, written production is still more costly than oral production. *International Journal of Psychology, 37*(4), 219–227. https://doi.org/10.1080/00207590244000070.

Burke, M. A., & Sass, T. R. (2013). Classroom peer effects and student achievement. *Journal of Labor Economics, 31*, 51–82. https://doi.org/10.1086/666653.

Calfee, R. C., & Miller, R. G. (2007). Best practices in writing assessment. In S. Graham, C. A. MacArthur, & J. Fitzgerald (Eds.), *Best practices in writing instruction* (pp. 265–286). Guilford.

Christensen, C. A. (2005). The role of orthographic-motor integration in the production of creative and well-structured written text for students in secondary school. *Educational Psychology, 25*(5), 441–453. https://doi.org/10.1080/01443410500042076.

Connelly, V., Gee, D., & Walsh, E. (2007). A comparison of keyboarded and handwritten compositions and the relationship with transcription speed. *British Journal of Educational Psychology, 77*, 479–492. https://doi.org/10.1348/000709906X116768.

Datchuk, S. M., & Kubina, R. M. (2012). A review of teaching sentence-level writing skills to students with writing difficulties and learning disabilities. *Remedial and Special Education, 34*(3), 180–192. https://doi.org/10.1177/0741932512448254.

De La Paz, S. (2007). Managing cognitive demands for writing: Comparing the effects of instructional components in strategy instruction. *Reading & Writing Quarterly, 23*(3), 249–266. https://doi.org/10.1080/10573560701277609.

De La Paz, S., Swanson, P. N., & Graham, S. (1998). The contribution of executive control to the revising by students with writing and learning difficulties. *Journal of Educational Psychology, 90*, 448–460. https://doi.org/10.1037/0022-0663.90.3.448.

Dockrell, J., Connelly, V., Walter, K., & Critten, S. (2018). The role of curriculum based measures in assessing writing products. In I. B. Miller, P. McCardle, & V. Connelly (Eds.), *Writing development in struggling learners* (pp. 182–197). Brill.

Englert, C. S., Stewart, S. R., & Hiebert, E. H. (1988). Young writers' use of text structure in expository text generation. *Journal of Educational Psychology, 80*(2), 143–151. https://doi.org/10.1037/0022-0663.80.2.143.

Feng, L., Lindner, A., Ji, S. R., & Joshi, R. M. (2019). The roles of handwriting and keyboarding in writing: A meta-analytic review. *Reading and Writing, 32*, 33–63. https://doi.org/10.1007/s11145-017-9749-x.

Fitzgerald, J., & Markham, L. (1987). Teaching children about revision in writing. *Cognition and Instruction, 4*, 3–24. https://doi.org/10.1207/s1532690xci0401_1.

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 of elementary school students. *School Psychology Review, 35*, 435–450.

Gersten, R., Jayanthi, M., Newman-Gonchar, R., Anderson, D., Spallone, S., & Taylor, M. J. (2020). *The reliability and consequential validity of two teacher-administered student mathematics diagnostic assessments (REL 2020–039)*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast. http://ies.ed.gov/ncee/edlabs.

Gersten, R., Keating, T., & Irvin, L. K. (1995). The burden of proof: Validity as improvement of instructional practice. *Exceptional Children, 61*, 510–519. https://doi. org/10.1177/001440299506100602.

Goldberg, A., Russell, M., & Cook, A. (2003). The effect of computers on student writing: A metaanalysis of studies from 1992 to 2002. *The Journal of Technology, Learning, and Assessment, 2*, 3–51.

Goodwin, A. P., & Ahn, S. (2010). A meta-analysis of morphological interventions: Effects on literacy achievement of children with literacy difficulties. *Annals of Dyslexia, 60*, 183–208. https://doi.org/10.1007/s11881-010-0041-x.

Graham, S. (2018). A revised writer(s)-within-community model of writing. *Educational Psychologist, 53*, 258–279. https://doi.org/10.1080/00461520.2018. 1481406.

Graham, S., Berninger, V. W., Abbott, R., Abbott, S., & Whitaker, D. (1997). Role of mechanics in composting of elementary school students: A new methodological approach. *Journal of Educational Psychology, 89*(1), 170–182. https://doi.org/10.1037/0022-0663.89.1.170.

Graham, S., Bruch, J., Fitzgerald, J., Friedrich, L., Furgeson, J., Greene, K., Kim, J., Lyskawa, J., Olson, C. B., & Smither Wulsin, C. (2016). *Teaching secondary students to write effectively (NCEE 2017-4002)*. National Center for Education Evaluation and Regional Assistance (NCEE), Institute of Education Sciences, U.S. Department of Education. Retrieved from the NCEE website: http://whatworks.ed.gov.

Graham, S., Capizzi, A., Harris, K. R., Hebert, M., & Morphy, P. (2014). Teaching writing to middle school students: A national survey. *Reading and Writing, 27*(6), 1015–1042. https://doi.org/10.1007/s11145-013-9495-7.

Graham, S., & Harris, K. R. (2017). Reading and writing connections: How writing can build better readers (and vice versa). In C. Ng, & B. Bartlett (Eds.), *Improving Reading and Reading engagement in the 21st century* (pp. 333–350). Springer Nature. https://doi.org/10.1007/978-981-10-4331-4_15.

Graham, S., Hebert, M., & Harris, K. R. (2011). Throw'em out or make'em better? State and district high-stakes writing assessments. *Focus on Exceptional Children, 44* (1). https://doi.org/10.17161/foec.v44i1.6913.

Graham, S., Kiuhara, S. A., & MacKay, M. (2020). The effects of writing on learning in science, social studies, and mathematics: A meta-analysis. *Review of Educational Research, 90*(2), 139–178. https://doi.org/10.3102/0034654320914744.

Graham, S., MacArthur, C. A., & Schwartz, S. (1995). Effects of goal setting and procedural facilitation on the revising behavior and writing performance of students with writing and learning problems. *Journal of Educational Psychology, 87*, 230–240. https://doi.org/10.1037/0022-0663.87.2.230.

Graham, S., McKeown, D., Kiuhara, S., & Harris, K. R. (2012). A meta-analysis of writing instruction for students in the elementary grades. *Journal of*

Educational Psychology, 104(4), 879–896. https://doi.org/10.1037/a0029185.

Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology, 99*(3), 445–476. https://doi.org/10. 1037/0022-0663.99.3.445.

Graham, S., & Santangelo, T. (2014). Does spelling instruction make students better spellers, readers, and writers? A meta-analytic review. *Reading and Writing, 27*, 1703–1743. https://doi.org/10.1007/s11145-014-9517-0.

Hammill, D. D., & Larsen, S. C. (2009). *Test of written language* (4th ed.). Pearson.

Hebert, M., Bohaty, J. J., Nelson, J. R., & Lambert, M. C. (2018). Identifying and discriminating expository text structures: An experiment with 4th and 5th grade struggling readers. *Reading and Writing, 31*(9), 2115–2145. https://doi.org/10.1007/s11145-018-9826-9.

Institute for Education Sciences, Technical Working Group. (2017). Future directions for writing research at the secondary level. Author. Retrieved from https://ies. ed.gov/ncer/whatsnew/techworkinggroup/.

Jones, D., & Christensen, C. A. (1999). The relationship between automaticity in handwriting and students' ability to generate written text. *Journal of Educational Psychology, 91*, 44–49. https://doi.org/10.1037/0022-0663.91.1.44.

Kellogg, R. T., Whiteford, A. P., Turner, C. E., Cahill, M., & Mertens, A. (2013). Working memory in written composition: An evaluation of the 1996 model. *Journal of Writing Research, 5*(2), 159–190.

Kent, S. C., & Wanzek, J. (2016). The relationship between component skills and writing quality and production across developmental levels: A meta-analysis of the last 25 years. *Review of Educational Research, 86*, 570–601. https://doi.org/10.3102/0034654315619491.

Kim, Y.-S. G. (2020). Structural relations of language and cognitive skills, and topic knowledge to written composition: A test of the direct and indirect effects model of writing. *British Journal of Educational Psychology, 90*(4), 910–932. https://doi.org/10.1111/bjep.12330.

Kim, Y.-S. G., Al Otaiba, S., Wanzek, J., & Gatlin, B. (2015). Towards an understanding of dimensions, predictors, and gender gap in written composition. *Journal of Educational Psychology, 107*, 79–95. https://doi.org/10.1037/a0037210.

Kim, Y.-S. G., & Schatschneider, C. (2017). Expanding the developmental models of writing: A direct and indirect effects model of developmental writing (DIEW). *Journal of Educational Psychology, 109*, 35–50. https://doi.org/10.1037/edu0000129.

Koutsoftas, A. D. (2016). Writing process products in intermediate-grade children with and without language-based learning disabilities. *Journal of Speech,* *Language, and Hearing Research, 59*, 1471–1483. https://doi.org/10.1044/2016_jslhr-l-15-0133.

Koutsoftas, A. D. (2018). Writing-process products of fourth- and sixth-grade children: A descriptive study. *Elementary School Journal, 118*, 632–653. https://doi.org/ 10.1086/697510.

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. https://doi.org/10.1044/0161-1461(2012/11-0018).

Koutsoftas, A. D., & Petersen, V. (2017). Written cohesion in children with and without language learning disabilities. *International Journal of Language and Communication Disorders, 52*, 612–625. https://doi.org/10.1111/1460-6984.12306.

Lee, Y. W., Gentile, C., & Kantor, R. (2008). *Analytic scoring of TOEFL CBT essays: Scores from humans and E-rater*. Educational Testing Service.

Lienemann, T. O., Graham, S., Leader-Janssen, B., & Reid, R. (2006). Improving the writing performance of struggling writers in second grade. *Journal of Special Education, 40*, 66–78. https://doi.org/10.1177/00224669060400020301.

Limpo, T., Alves, R. A., & Connelly, V. (2017). Examining the transcription-writing link: Effects of handwriting fluency and spelling accuracy on writing performance via planning and translating in middle grades.

Learning and Individual Differences, 53, 26–36. https://doi.org/10.1016/j.lindif.2016.11.004.

McCardle, P., Miller, B., & Connelly, V. (2018). Approaches to improving writing research, instruction, and performance. In B. Miller, P. McCardle, & V. Connelly (Eds.), *Writing development in struggling learners* (pp. 201–215). Brill.

McCoach, D. B., & Adelson, J. L. (2010). Dealing with dependence (part I): Understanding the effects of clustered data. *The Gifted Child Quarterly, 54*(2), 152–155. McCutchen, D. (2011). From novice to expert: Implications of language skills and writing-relevant knowledge for memory during the development of writing skill. *Journal of Writing Research, 3*(1), 51–68. https://doi.org/10.17239/jowr-2011.03.01.3.

McKeown, D., Brindle, M., Harris, K. R., Sandmel, K., Steinbrecher, T. D., Graham, S., … Oakes, W. P. (2019). Teachers' voices: Perceptions of effective professional development and classwide implementation of self-regulated strategy development in writing. *American Educational Research Journal, 56*, 753–791. https://doi. org/10.3102/0002831218804146.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2009). Linguistic features of writing quality. *Written Communication, 27*, 57–86. https://doi.org/10.1177/ 0741088309351547.

McNamara, D. S., Louwerse, M. M., Cai, Z., & Graesser, A. (2005). Coh-Metrix version 3.0. Retrieved from http://cohmetrix.memphis.edu.

McNaughton, D., Hughes, C., & Ofiesh, N. (1997). Proofreading for students with learning disabilities: Integrating computer and strategy use. *Learning Disabilities Research & Practice, 12*, 16–28. https://doi.org/10.1177/002221949703000608.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.

Messick, S. (1998). Consequences of test interpretation and use: The fusion of validity and values in psychological assessment. *ETS Research Report Series, 1998*(2), i–32. https://doi.org/10.1002/j.23338504.1998.tb01797.x.

Michigan Department of Education. (2017). Michigan Student Test of Educational Progress (MSTEP) guide to reports. Retrieved from https://www.michigan.gov/documents/mde/2017_MSTEP_GTR_598970_7.pdf.

Mo, Y., & Troia, G. A. (2017). Similarities and differences in constructs represented by U.S. states' middle school writing tests and the 2007 National Assessment of educational Progress writing assessment. *Assessing Writing, 33*, 48–67.

Monroe, M., & Sherman, E. (1966). *Group diagnostic reading aptitude and achievement tests*. C.H. Nevins Printing.

Muthen, L. K., & Muthen, B. O. (2019). *Mplus 8.3*. Author.

National Assessment of Educational Progress (NAEP), U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. (2011). Writing assessment. Available at www.nationsreportcard.gov/writing_2011/.

National Center on Intensive Instruction (NCII). (2019). Academic progress monitoring tools chart rating rubric. Retrieved from https://intensiveintervention.org/sites/default/files/NCII_AcadProgMonitoring_RatingRubric_Aug2019.pdf.

National Commission on Writing. (2004). *Writing: A ticket to work or a ticket out: A survey of business leaders*. College Board.

National Governors Association & Council of Chief School Officers. (2010). Common core state standards. Available from www.corestandards.org/.

National Research Council (NRC). (2001). *Knowing what students know: The science and design of educational assessment*. Committee on the Foundations of Assessment.

Olinghouse, N. G. (2008). Student- and instruction-level predictors of narrative writing in third-grade students. *Reading and Writing: An Interdisciplinary Journal, 21*, 3–26. https://doi.org/10.1007/s11145-007-9062-1.

Olive, T. (2014). Toward a parallel and cascading model of the writing system: A review of research on writing processes coordination. *Journal of Writing Research, 6* (2), 173–194. https://doi.org/10.17239/jowr-2014.06.02.4.

O'Rourke, L., Connelly, V., & Barnett, A. (2018). Understanding writing difficulties through a model of the cognitive processes involved in writing. In B. Miller, P. McCardle, & V. Connelly (Eds.), *Writing development in struggling learners* (pp. 11–28). Brill.

Puranik, C. S., & Al Otaiba, S. (2012). Examining the contribution of handwriting and spelling to written expression in kindergarten children. *Reading and Writing: An Interdisciplinary Journal, 25*, 1523–1546. https://doi.org/10.1007/s11145-011-9331-x.

Reifman, A., & Keyton, K. (2010). Winsorize. In N. J. Salkind (Ed.), *Encyclopedia of research design* (pp. 1636–1638). Sage.

Rouse, A. G., & Kiuhara, S. A. (2017). SRSD in writing and professional development for teachers: Practice and promise for elementary and middle school students with learning disabilities. *Learning Disabilities Research & Practice, 32*, 180–188. https://doi.org/10.1111/ldrp.12140.

Saddler, B., & Asaro, K. (2007). Increasing story quality through planning and revising: Effects on young writers with learning disabilities. *Learning Disability Quarterly, 30*, 223–234. https://doi.org/10.2307/25474635.

Santangelo, T., & Graham, S. (2016). A comprehensive meta-analysis of handwriting instruction. *Educational Psychology Review, 28*(2), 225–265. https://doi.org/10. 1007/s10648-015-9335-1.

Santangelo, T., & Olinghouse, N. G. (2009). Effective writing instruction for students who experience writing difficulties. *Focus on Exceptional Children, 42*(4), 1–20.

Silverman, R. D., Coker, D., Proctor, C. P., Harring, J., Piantedosi, K., & Hartranft, A. M. (2015). The relationship between language skills and writing outcomes for linguistically diverse students in upper elementary school. *Elementary School Journal, 116*, 103–125. https://doi.org/10.1086/683135.

StataCorp. (2013). *Stata version 13.0*. StataCorp LP.

Torrance, M., Arrimada, M., & Gardner, S. (2020). Child-level factors affecting rate of learning to write in first grade. *British Journal of Educational Psychology, 91*,

714–734.

Tracy, B., Reid, R., & Graham, S. (2009). Teaching young students strategies for planning and drafting stories: The impact of self-regulated strategy development. *Journal of Educational Research, 102*, 323–331. https://doi.org/10.3200/joer.102.5.323-332.

Troia, G. A., Brehmer, J., Glause, K., Reichmuth, H., & Lawrence, F. R. (2020). Direct and indirect effects of literacy skills and writing fluency on writing quality across three genres. *Education Sciences, 10*, 297. https://doi.org/10.3390/educsci10110297.

Troia, G. A., & Graham, S. (2002). The effectiveness of a highly explicit, teacher-directed strategy instruction routine: Changing the writing performance of students with learning disabilities. *Journal of Learning Disabilities, 35*, 290–305. https://doi.org/10.1177/00222194020350040101.

Troia, G. A., Graham, S., & Harris, K. R. (1999). Teaching students with learning disabilities to mindfully plan when writing. *Exceptional Children, 65*, 235–252. https://doi.org/10.1177/001440299906500208.

Troia, G. A., Harbaugh, A. G., Shankland, R. K., Wolbers, K. A., & Lawrence, A. M. (2013). Relationships between writing motivation, writing activity, and writing performance: Effects of grade, sex, and ability. *Reading and Writing: An Interdisciplinary Journal, 26*, 17–44. https://doi.org/10.1007/s11145-012-9379-2.

Troia, G. A., Shen, M., & Brandon, D. L. (2019). Multidimensional levels of language writing measures in grades four to six. *Written Communication, 36*(2), 231–266. https://doi.org/10.1177/0741088318819473.

Truckenmiller, A. J., McKindles, J. V., Petscher, Y., Eckert, T. L., & Tock, J. L. (2020). Expanding curriculum-based measurement in written expression for middle school. *Journal of Special Education, 54*, 133–145. https://doi.org/10.1177/0022466919887.

Truckenmiller, A. J., & Petscher, Y. (2020). The role of academic language in written composition in elementary and middle school. *Reading and Writing, 33*(1), 45–66. https://doi.org/10.1007/s11145-019-09938-7.

Truckenmiller, A. J., Shen, M., & Sweet, L. E. (2021). The role of vocabulary and syntax in informational written composition in middle school. *Reading and Writing, 34*, 911–943. https://doi.org/10.1007/s11145-020-10099-1.

Tukey, J. (1977). Some thoughts on clinical trials, especially problems of multiplicity. *Science, 198*(4318), 679–684. https://doi.org/10.1126/science.333584.

Wagner, R. K., Puranik, C. S., Foorman, B., Foster, E., Wilson, L. G., Tschinkel, E., & Kantor, P. T. (2011). Modeling the development of written language. *Reading and Writing, 24*(2), 203–220. https://doi.org/10.1007/s11145-010-9266-7.

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 56–75). Sage Publications, Inc.

Wilson, J., Roscoe, R., & Ahmed, Y. (2017). Automated formative writing assessment using a levels of language approach. *Assessing Writing, 34*, 16–36. https://doi. org/10.1016/j.asw.2017.08.002.

Wright, T. S., & Cervetti, G. N. (2017). A systematic review of the research on vocabulary instruction that impacts text comprehension. *Reading Research Quarterly, 52*, 203–226. https://doi.org/10.1002/rrq.163.