ASSESSING STUDENTS' SKILLS AT WRITING ANALYTICALLY IN RESPONSE TO TEXTS

ABSTRACT

Despite the importance of writing analytically in response to texts, there are few assessments measuring students' mastery of this skill. This manuscript describes the development of a response-to-text assessment (RTA) intended for use in research. In a subsequent validity investigation we examined whether the RTA distinguished among classrooms in students' ability to write analytically in response to text and whether measures of teaching predicted this variation. We demonstrate that the RTA was correlated with the state standardized assessment, but did not overlap with this accountability test completely and, additionally, that more variation between classrooms existed on the RTA. Students' opportunities for reasoning and extended writing in the classroom were significantly associated with RTA scores. The findings suggest that the RTA can be a valuable tool for conducting research on students' attainment of analytic writing skills and for understanding how teaching relates to student achievement on these skills.

Richard Correnti
Lindsay Clare Matsumura
UNIVERSITY OF PITTSBURGH

Laura Hamilton
RAND CORPORATION

Elaine Wang
UNIVERSITY OF PITTSBURGH

ANALYTIC writing in response to text—that is, the ability to interpret and evaluate texts, construct logical arguments based on substantive claims, and marshal appropriate evidence in support of these claims—is fundamental to academic success (Common Core State Standards Initiative, 2010; National Commission on Writing, 2003). Students who lack or only partially master the ability to analyze texts often struggle in the secondary grades and/or are

THE ELEMENTARY SCHOOL JOURNAL VOLUME 114, NUMBER 2

likely to find college-level coursework too difficult to complete (Allensworth, Correa, & Ponisciak, 2008). Despite the importance of writing analytically in response to texts, little systematic information about students' progress toward attaining this skill is available to most schools and districts. Instead, the most readily available information about students' literacy skills is generally obtained from large-scale state achievement tests and other standardized tests that typically represent reading comprehension and writing as separate skills. Reading comprehension is usually measured through a series of multiple-choice questions and brief constructed responses based on texts that are not explicitly connected to the academic content to which students are exposed in the school curriculum. Hirsch and Pondiscio (2010) refer to these as "content-free" assessments, and they note that such assessments fail to recognize that comprehension of text is inherently content laden. Writing, meanwhile, is often excluded from large-scale testing programs. When included, writing typically involves having students respond to an open-ended prompt that is not connected to a rich text (Jeffery, 2009). Despite the fact that the bulk of challenging academic work in many content areas lies at the intersection of reading comprehension and writing, few large-scale assessments integrate these two domains in the context of rich textual material.

In this article, we describe our research focused on developing an assessment that fills this gap; that is, one that measures students' ability to write analytically in response to text. Our approach to assessment sampled from the target domains of reading comprehension and writing to create an assessment that required students to reason about the text and then write an extended response, but did not require students to read the text independently. While noting this critical distinction between our assessment and similar response-to-literature formats requiring students to first read the text independently, we refer to this as a *response-to-text assessment* (RTA) because students were expected to analyze the content of the text in their written responses.

Measures such as the RTA could potentially serve many purposes. For example, measuring students' ability to write analytically, incorporating content from a text or texts could draw attention to this academic skill and thereby encourage educators to emphasize it. In addition, assessments that provide teachers information about students' attainment of analytic text-based writing skills could potentially help teachers evaluate their own instructional approach and encourage them to engage in teaching practices to develop those skills in students. Our particular goal in developing the RTA, however, is to create a measure that can be used to further educational research. It is this purpose that we address in this study.

Much has been written about the potential for "higher-order" teaching to influence students' learning outcomes (see, e.g., Abrami et al., 2008, and Nickerson, 1989, for reviews of this work). Teaching of complex skills is thought to enhance students' ability to achieve on measures of rote learning as well as assessments meant to demonstrate reasoning or high cognitive demand skills such as critical thinking (e.g., Abrami et al., 2008). While it is generally accepted that higher-order instruction varies between classrooms (Knapp, 1995; Raudenbush, Rowan, & Cheong, 1993), there is a corresponding lack of assessments that demonstrate the effects of these instructional differences on student learning. Yet, assessments that gauge the extent to which students are capable of more complex reasoning are important for understanding whether complex learning is being achieved in all classrooms (Resnick, 1987, 2010) and whether certain approaches to teaching are associated with that

learning (Bereiter & Scardamalia, 1987). This information is critical to building a knowledge base of the teaching profession that can inform professional education and guide the development of more effective instructional interventions and policies (Hiebert, Gallimore, & Stigler, 2002).

We were especially motivated to create a response-to-text assessment in the context of the current policy environment in which states are adopting the Common Core State Standards (CCSSI, 2010). A key feature of the Common Core State Standards is the emphasis on students' engagement with complex texts and, beginning in the upper elementary grades, ability to draw inferences from a given text and use evidence to support their assertions (CCSSI, 2010). Our hunch is that assessments measuring students' ability to write analytically in response to a text more so than general tests of achievement will be more sensitive to natural variation in instruction. A similar argument was made in at least one prominent policy study, where the outcome—a nontraditional student assessment—was sensitive to covariates distinguishing between different professional development experiences (Cohen & Hill, 2001).

While two consortia are currently developing the state-level assessments linked to the CCSS, we believe that additional measures that are not part of the state's accountability system will be needed in research. We anticipate that instructional interventions will increasingly focus on developing teachers' ability to teach students to respond analytically to texts in writing; correspondingly, measures of student learning aligned with these instructional processes will be needed. Lack of alignment between intervention goals and assessed student learning outcomes is a persistent problem in education research (e.g., Berends, Bodilly, & Kirby, 2002; Correnti & Rowan, 2007; Stebbins, St. Pierre, Proper, Anderson, & Cerva, 1977). Effects of interventions aimed at increasing teachers' ability to teach complex academic skills may not be demonstrated when students' learning outcomes are assessed on tests that measure lower-level skills. When assessments are not well aligned with interventions, it is impossible to know if the interventions did not show desired effects because they were not successful, or because the intended outcomes were not adequately measured (Caro, 1971; Raudenbush, 2007). In the following section, we elaborate on why a focus on overlap between what is taught and what is tested is so important for identifying links between teaching and learning.

Properties of Overlap and Specificity

Numerous researchers have demonstrated that the alignment of assessments with the object of study has ramifications for the ability to find effects of interventions or professional development (PD) programs. A special case of alignment arises when researchers attempt to study teaching effects on student learning. Here, researchers have demonstrated how overlap between students' opportunities to learn and the tested content has an influence on research findings (Barr & Dreeben, 1983; Berliner, 1981; D'Agostino, Welsh, & Corson, 2007; Ruiz-Primo, Shavelson, Hamilton, & Klein, 2002). Assessments can be biased if the overlap between different curricula and assessments varies (Leinhardt & Seewald, 1981).

Overlap was a central motivation for cross-national studies such as the Third International Mathematics and Science Study (TIMSS; see, e.g., Schmidt et al., 2001), but the value of the concept would also seem to extend to classroom-level studies of

teaching. That is, careful alignment between teaching and measures of learning is more likely to provide a "fairer" test for finding associations between the two, since researchers can avoid the specification error inherited when trying to associate (specific) measures of teaching with more general measures of student learning. Moreover, students are likely to perform better when they have had repeated exposure to both the *content* and the *format* of the assessment, since what is taught and what is tested both represent sampled domains (Cooley & Leinhardt, 1980).

On this latter point, the specificity[1] or distinctiveness of the assessed domain is also likely to have bearing on the likelihood of finding an association. By narrowing assessments to specific (and potentially infrequently measured) skills, researchers should be more likely to identify teaching-learning associations, provided the measures of teaching have good overlap with the assessments. These specific associations are important to begin to identify since they are the bedrock of a professional knowledge base of teaching. It is especially instructive to think of these specific associations in relation to the accountability test, since they may highlight where learning that extends beyond the skills generally tested in large-scale assessments is occurring in classrooms.

These twin properties of overlap and specificity help motivate our validity investigation. Our assessment contains dimensions of both specificity and overlap. On the RTA we have tried to assess student ability to respond analytically to a text in an extended written response, a skill that is complex but also fairly distinct when compared with the broad range of literacy skills typically measured on large-scale assessments. Furthermore, it is possible to imagine measures of teaching with good overlap to the assessment, since these measures of teaching are also infrequently captured and since students' skills in analytic writing are not likely to develop without practice (i.e., without regular exposure to extended writing and text analysis opportunities in their literacy instruction). In the context of a unified view of validity (Kane, 2006; Messick, 1994), we therefore hypothesize that due to our focus on reasoning and writing we are well positioned to find theoretically relevant associations between our specific measures of teaching and learning. This evidence will inform our understanding of the construct validity of inferences made from the RTA.

Claims

This article examines the validity of the RTA for the purpose of documenting student learning in research contexts. An investigation of validity typically involves the collection of a wide range of evidence that provides a scientific basis for a specific score interpretation (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Kane, 2006). Most assessments are designed to serve specific purposes, and each purpose involves an interpretation of scores that should be subjected to a validity investigation. As described earlier, our primary purpose in developing the RTA was to create a measure that can be used in studies that require evidence about students' ability to write meaningfully in response to the content of a text and about the instructional processes that promote that ability. To be useful for research, this measure would need to provide information about differences in students' writing skills across classrooms and be sensitive to the effects of instructional practices that are intended to enhance attainment of these skills. Therefore, we investigate the following claims

with our data: (1) the RTA distinguished among classrooms in students' ability to write in response to text after adjusting for student demographic differences and prior achievement; (2) the RTA is more aligned to teaching practices requiring student reasoning and extended writing than typical forms of assessment (i.e., standardized state tests). The RTA, therefore, is positioned to demonstrate that differences between classrooms in teaching reasoning and extended writing result in differences in student learning outcomes. In the sections that follow, we examine results from student performance on the two assessments—the Maryland School Assessment (MSA) and the RTA—that we used to investigate these claims.

Development of the Response-to-Text Assessment (RTA)

The development of the RTA was guided by three central principles. First, we wanted the assessment to build on an authentic text and support students to reason about the text and cite supporting evidence. Second, given that the emphasis was on measurement of students' skills at reasoning and writing, we designed our administration procedure to eliminate other potential sources of measurement error for assessing those skills. Third, we sought to make our assessment relevant within the current research context. Thus, we sought to align the RTA with the Common Core State Standards (CCSSI, 2010) because as schools and districts respond to state adoption of the standards, they are likely to drive choices regarding curriculum and instruction.

We had three criteria for choosing the texts on which the assessment was based. First, we wanted the text to be authentic (the kind of text students might encounter in their own classroom) but brief. Brevity was important because in addition to having the text read to them, we wanted students to have time to reason about the text and also write an extended response. Second, the text content had to support the development of an open-ended prompt that would invite multiple interpretations and for which students could provide textual evidence to support their assertions (Beck & McKeown, 2001). Finally, in keeping with the spirit of the Common Core State Standards, we wanted the text to be grade-level appropriate but also challenging. The latter criterion is the most problematic to satisfy given the difficulty of anticipating the reading levels of the students we intended to assess. Nevertheless, post-hoc analysis of the lexile level of each text convinced us that our text selection struck a middle ground between our pilot teachers' concerns over readability of the text and the ideals of the Common Core State Standards stating that texts should challenge readers.[2]

In an attempt to minimize measurement error from unwanted sources, we made several decisions to standardize the administration of the RTA. Since our assessment was designed to detect differences among students in their ability to respond in writing to a text, we did not want reading fluency to confound the measurement of students' written responses. Therefore, we decided to have the teacher read the text while the students followed along with their own copy of the text. After this initial reading, students could refer back to the text at any time. We also sought to minimize the role that vocabulary knowledge played in the measurement of student writing ability; therefore, we used call-out boxes to define several vocabulary words (e.g., *hasty*, *deranged*, and *irrigation*).

Finally, based on research showing the important role of text discussions for supporting students' comprehension (Murphy, Wilkinson, Soter, Hennessey, & Al-

exander, 2009; Nystrand, 2006), we identified points in the texts at which teachers would stop to ask standardized questions. Drawing on *Questioning the Author* (Beck & McKeown, 2006), we developed these standardized questions to aid students' literal comprehension of the text as it was being read. Our protocol (much like a standard interview protocol) included initial questions along with expected student responses and standardized follow-up questions if students did not respond as expected. These questions were designed to give all students access to a literal understanding of important story elements so they all had the potential to address the prompt. These decisions were motivated by our desire to measure how students analyzed the content of the text beyond a literal understanding and then communicated their analysis in writing.

Previous Study of the RTA

We have engaged in a cycle of iterative improvements to the RTA (Diamond & Powell, 2011). In each attempt we have reconsidered our prompts as well as our administration and scoring procedures based on feedback received from literacy experts, teachers, and pilot studies. Our initial pilot of the RTA was conducted in 30 classrooms in western Pennsylvania. We administered prompts from four different texts (two were fiction and two were nonfiction) in a fully crossed design where each prompt was randomly assigned to equal numbers of grade 4 and 5 classrooms.

Three findings from the pilot informed our current work. First, for 12 of the classrooms we had scores for the same students on the Pennsylvania System of School Assessments (PSSA) reading test and the RTA. We found more variance between classrooms on the RTA (18%) compared to the PSSA (12%) after adjusting for background characteristics. Second, using a teacher survey, we found that variation in teachers' self-reports of the extent to which they had students integrate writing with their comprehension instruction[3] was associated with student performance on the RTA. The effect size for this covariate was .35, indicating that one unit higher on this item (which was on a five-point scale) resulted in higher classroom average RTA scores of a little more than one-third of a standard deviation. Third, in a hierarchical linear model adjusting for factors such as prompt difficulty and student characteristics, students in grade 5 scored higher than fourth graders by about .4 standard deviations on the assessments. This confirmed that scores increased with developmental age and/or school experiences, suggesting the RTA measures were sensitive enough to capture gross differences in ability resulting from maturation and exposure to the grade 5 curriculum. Based on this early pilot, we selected two of the four prompts for further revision—one fictional to be administered to fourth graders and one nonfictional to be administered to fifth and sixth graders. This iteration of the RTA forms the basis for our current study.

Current Version of the RTA

The grade 4 form of the current version of the RTA is based on a short story by James Marshall titled "Rats on the Roof." The story involves a pair of dogs who have a rat problem and enlist a snobbish cat to help them. The cat solves the problem but, ironically, not in the way the dogs intended. The prompt asks students, "Is the Tomcat someone you would want to help you with a problem? Why or why not? Use

at least 3 to 4 examples from the text to explain your answer." We based the grade 5 form on a feature article from *Time for Kids* about a United Nations–supported effort to eradicate poverty in a rural village in Kenya through the Millennium Villages project. The prompt asks students, "Why do you think the author thinks it's important for kids in the United States to learn about what life was like in Kenya before and after the Millennium Villages project? Make sure to include at least 3 examples of what life in Kenya was like before the Millennium Villages project and what life is like now." Each of these prompts was intended to direct students toward formulating an organized piece of writing that insightfully analyzed an aspect of the given text. Furthermore, the prompts were designed to encourage students to elaborate upon their responses and to integrate multiple instances of appropriate textual evidence. In short, we sought to create a task that embodied the principles of a "thinking curriculum" (Resnick & Klopfer, 1989) and of "thoughtful literacy" (Brown, 1991).

We developed multiple rating categories to evaluate not only students' thinking about the text, but their skill at marshaling evidence in support of their claims as well as other criteria associated with effective analytic writing. Our goal for creating multiple rating categories was to provide detailed information about students' analytic writing skills, as well as facilitate the investigation of hypotheses related to how instructional behaviors might influence different aspects of students' writing. The process included iterative revisions to the rubric itself (see App. A). In constructing the scoring criteria, we also sought to ensure that the language and expectations reflected those of the Common Core State Standards.[4] Over iterative sessions, a four-person group calibrated against the rubric to ensure uniform understanding and application of the criteria. Interrater reliability is reported below.

Each of the five criteria was rated on a four-point scale (1 = low to 4 = excellent). *Analysis* assessed students' ability to demonstrate a clear understanding of the purpose of the literary work and to make valid and perceptive conclusions that inform an insightful response to the prompt. *Evidence* captured the degree to which students select and use details, including direct quotations from the text to support their key idea. *Organization* assessed the degree to which students' responses exhibit a strong sense of beginning, middle, and end, and demonstrate logical flow between sentences and ideas. The *Style* criterion awarded points for varied sentence lengths and complex syntactical structures, multiple uses of tier 2 vocabulary (e.g., words like *instructions*, *fortunate*, *miserable*, *appreciate*), and correct application of sophisticated connectives (e.g., *however*, *meanwhile*). Finally, students' scores on *Mechanics/Usage/Grammar/Spelling (MUGS)* reflected their ability to adhere to grade-appropriate standard writing conventions.

Interrater reliability was calculated on about 20% of the sample. A graduate student rater (with prior experience scoring essays on rubrics) coded all 426 pieces of student writing. To calculate interrater reliability, the study's project manager coded 86 pieces of student writing randomly chosen from the larger sample, including 45 responses to the "Rats on the Roof" prompt and 41 responses to the "Millennium Villages" prompt. We examined a crosstab of the two raters. It showed that the exact match between raters was 79%, with only two instances of raters differing by more than one. A Cohen's kappa was also calculated, which reports the exact match agreement adjusted for the distribution of scores across categories. Cohen's kappa (.672, $\chi^2 = 603.94$, $df = 9$) and the Pearson correlation ($r = .828$) both indicate moderately high agreement between raters overall.

We further examined interrater agreement by each of the five rating criteria and for each of the two prompts. Results from the five criteria showed that exact agreement ranged from a low of 70% (Evidence) to a high of 88% (Style). Cohen's kappa values for the five criteria from lowest to highest were Evidence (.55), Analysis (.62), MUGS (.68), Organization (.69), and Style (.81). While agreement rates varied, even the lowest category (Evidence) had a Pearson's correlation of .80 between raters and only a single instance of raters being off by more than one. Agreement rates for exact matches for "Rats on the Roof" (77%) and "Millennium Villages" (78%), respectively, were quite similar.[5]

Current Study Data

Sample

The data used to investigate our claims were collected in 18 classrooms from a single urban district in Maryland. Data on students were collected in the spring, including measures of learning on two different assessments. First, we examined the reading portion of the Maryland School Assessment (MSA), which is the state standardized test that all students from grades 3–8 must take. It was administered over 2 days in March, with about 90 minutes of testing time each day. The test consisted of 33 multiple-choice questions on vocabulary, word study, and reading comprehension, and four brief constructed responses (BCR). A sample BCR prompt asks the following: "Explain how the setting affects the actions of the characters in this story. In your response, use details from the story that support your explanation" (Maryland State Department of Education [MSDE], 2012). Students must respond to each prompt within the eight lines provided. Accommodations for students with individualized education plans or 504 plans can be provided upon application to and approval from the Department of Education. Limited English proficient (LEP) students who have been enrolled for more than 1 year must also take the MSA. Again, accommodations may be approved. In all cases, students completed the assessment individually and independently of any teacher input. In terms of scoring, the BCR is given a rating of 0–3, depending on the extent to which the response addresses the "demands of the question" and "uses test-relevant information to show understanding" (MSDE, 2012). The overall test score consisted of three subscales: General Reading (15 multiple-choice items), Literary Reading and Information Reading (nine multiple-choice and two BCRs each). The test publisher created scale scores for the test overall and for each subscale.

The RTA was administered in one 60-minute session by teachers during the last week of May. Teachers in eight grade 4 classrooms administered the prompt based on the fictional text "Rats on the Roof" and nine grade 5 teachers and one grade 6 teacher administered the prompt based on the "Millennium Villages" informational text. All students present on testing day in the participating classes took the RTA; no makeup for absent students was available. Students received the testing accommodations they normally receive through regular instruction. Teachers were directed to spend the first 15 minutes reading the text aloud while students followed along. Teachers were also asked to read predefined vocabulary and stop at designated points to ask and engage students in discussions of standardized prompts. Expected student responses and follow-up questions were provided to help guide student comprehension. At the end, teachers read the writing prompt with students and allowed them 45 minutes to respond on the two pages of lined paper provided. Students were encouraged to brainstorm and plan as needed.

Information on student demographic characteristics, including prior year achievement on the MSA in both reading and math, was collected from the district for all 426 students who took the RTA. The demographic characteristics of students in our sample were roughly representative of the larger district. About 56% of the students received free (45%) or reduced (11%) price lunch, 11% had an individualized education plan (IEP), and students were absent, on average, about 7.5 days per year. Students in our sample were predominantly minority, indicating the following group affiliations: Black (80%), Hispanic (12%), Native American (11%), Asian (5%), and White (3%). In our prediction models we adjusted for these student background characteristics.

These 426 students were nested in 18 classrooms. Seventy-eight percent of the teachers in the sample had already received their Master's (61%) or Ph.D. (17%) degree. Additionally, 18% had advanced professional certification. Teachers' years of teaching experience ranged from 2 to 38, with an average of 17. The teachers in each of the 18 classrooms also participated in our larger data-collection efforts. These efforts included an annual survey, daily literacy logs (30 daily surveys over the course of the year in the fall, winter, and spring), and six response-to-literature assignments (three each in winter and spring). Sixteen of the 18 teachers had complete data across the survey, logs, and assignments. The remaining two teachers had log and assignment data, but did not complete the annual survey. Rather than use listwise deletion to remove cases without complete data on instruction, we opted for a strategy of multiple imputation. Following procedures outlined in Peugh and Enders (2004), we imputed five data sets for the teacher survey measures used in our analyses under the assumption that data were missing at random.[6] To accomplish this, a two-level multiple imputation was conducted in MPlus6.12 (Muthén & Muthén, 2010) using the regression command.

Instructional Scales

Using the five imputed data sets, we created two instructional scales from items on the survey, logs, and assignments—one measure examined students' opportunities for reasoning and extended writing while the second utilized teachers' self-reports of their reading comprehension instruction. Both of these instructional factors were used in prediction models on the RTA and MSA.

Student reasoning and extended writing factor. Six covariates were combined to create a composite factor measuring students' opportunities for reasoning and extended writing. This covariate was hypothesized to be well aligned to the RTA because it measures the opportunity structure for students to develop analytical, text-based writing skills by sampling from four target domains: time devoted to writing, activities analyzing and/or synthesizing text(s) in the context of discussing or doing writing, opportunities for elaborated communication, and exposure to writing tasks of a high cognitive demand. These student experiences provide opportunities for students to practice skills aligned in content and format with the skills presumably measured by the RTA. Covariates measuring this opportunity structure are described in Table 1 and are further elaborated in Correnti, Matsumura, Hamilton, and Wang (2012).

The variables contributing to this composite include (1) the frequency teachers reported integrating writing into their text discussions, (2) a writing scale from a measurement model of log data, (3) a ratio of high-cognitive-demand items to all other items from the measurement model of log data (primarily measuring integration of comprehension and writing as reciprocal processes), (4) the cognitive de-

Table 1. Items Contributing to Measure of Students' Opportunities for Reasoning and Extended Writing[a]

Measures	Description	Range	Mean	*SD*
Surveys:				
Frequency of reasoning/writing integrated in text discussions	This covariate is a factor of 4 survey responses to an item stem asking the following: "Below are a set of items pertaining to classroom text discussions. Considering constraints on your teaching please indicate the frequency students engage in each element in your text discussions." Items were answered on the following scale: never, rarely, sometimes, often, and almost always. Four items formed a single factor explaining 63% of the variance in the items, and included, (1) students identify the author's purpose, (2) students discuss elements of the writer's craft, (3) students make connections between ideas/ literary elements within or across texts, and (4) students analyze and evaluate each other's assertions.	2.5–5	3.20	.75
Logs:[b]				
Writing	Scale of writing items (pre-writing, writing practice, literary techniques, revise, edit, share, teacher comment on writing, teacher-directed writing instruction, integrate comprehension, write connected paragraphs) across all logs contributed and weighted by time. This scale represents the exposure students had to writing across the year.	−1.47–2.24	0	1
Integration of comprehension with writing	Scale of 8 items (analyze/evaluate in comprehension, focus on writing same day as focus on comprehension, integrate writing with comprehension, provide extended answers in comprehension, literary techniques, integrate reading comprehension with writing, substantive revisions, wrote multiple paragraphs) contrasted with all other log items across all logs contributed and weighted by time. This scale represents a ratio of the frequency teachers decided to integrate comprehension and writing instruction relative to all other content.	−1.46–1.74	0	1
Assignments:				
Enacted cognitive demand rating	This dimension is assessed on a 4-point scale (1 = poor, 4 = excellent) and focuses on the degree to which an assignment supports students to apply higher-level, analytic thinking skills (as opposed to recalling or identifying basic information from a text) and use appropriate evidence and details from a text to support their assertions. Interrater exact agreement of 90%.	1.33–2.83	2.08	.39

Table 1. (*Continued*)

Measures	Description	Range	Mean	*SD*
Percent of teacher-created assignments	Percent of challenging assignments that went beyond brief constructed responses (restricting students to respond within a box) or curriculum-generated worksheets. Despite explicit instructions that read, "Please select a RESPONSE TO LITERATURE assignment you consider to be CHALLENGING for your students. This may include (but is not limited to): a summary of a text, an evaluation of a book, an analysis of a character, an essay comparing and contrasting texts," a majority of assignments received were directly from the district curricula.	0–100	32	.34
Average length of written response	The average number of words students wrote in response to challenging assignments, averaged across 6 assignments with 4 students per assignment (2 demonstrating high ability, 2 demonstrating average ability).	38–230	112.42	59.75

[a] Data reduction was accomplished through factor analysis using SPSS 19.0 specifying principal axis factoring and an oblique rotation. A single factor was obtained explaining just over 60% of the variance in the six items. Examination of the scree plot indicated a single factor was preferred.

[b] The frequency of days teachers reported using each of 28 dichotomous items in comprehension (10 items), writing (9 items), and word analysis (9 items) was reported in Correnti et al. (2012). These dichotomous items were nested within days, and days were nested in teachers to construct scales from a measurement model (e.g., see Raudenbush et al., 1991). The models weighted each occasion by the number of minutes teachers reported teaching that day. From these measurement models we obtained an empirical Bayes residual from instructional scales estimated simultaneously including writing and integration of comprehension with writing.

mand rating of the challenging assignment tasks we collected, (5) the proportion of challenging assignments that went beyond either worksheets or brief constructed responses of a paragraph or less, and (6) the average number of words students wrote in response to their assignments.[7] The single composite formed from the six covariates and described in Table 1 was normally distributed.

Comprehension factor. Three covariates from self-report measures on the survey and logs were used to create this composite. These covariates are further described in Table 2. The first covariate from the survey is calculated from the product of teacher-reported time spent in comprehension and teacher-reported time spent inferring the author's meaning or citing evidence from text to support assertions when doing comprehension text discussions. The second covariate from the survey represents the frequency with which students participated in routine text discussion activities, such as recounting or sequencing events in a story. The third covariate was developed from the logs and utilized a measurement model to form a scale of comprehension similar to the scale reported in Rowan, Camburn, and Correnti (2004). The single composite formed from these three covariates was normally distributed.

Analytic Methods

Our data analyses were constructed to seek evidence related to our earlier claims. The first claim sought to understand the extent to which the RTA distinguished among classrooms on students' ability to write in response to texts. Implicit in the claim is that the RTA distinguishes among classrooms such that we might draw valid infer-

Table 2. Items Contributing to Measure of Students' Opportunities for Comprehension[a]

Measures	Description	Range	Mean	*SD*
Surveys:				
Time spent in text discussions inferring and citing evidence	This covariate was calculated from teachers' self-report of two separate time estimates. First, teachers were asked, "Consider the total amount of time students spent on language arts in a typical week. What proportion of their time was spent on the following? (Proportion of time should total 100%)." Teachers reported the proportion of time they spent in each of four options including (1) reading text and improving reading skills, (2) writing, (3) assessing students' understanding/comprehension through multiple choice, fill-in-the-blank, or practice for the MSA, and (4) text discussion activities. We were interested in the latter since it is our most direct measure of teaching comprehension of the text. Second, teachers were asked, "Consider the total amount of time students spent in text discussions. What proportion of time was spent on the following? (Proportion of time should total 100%)." Teachers reported the proportion of time they spent in each of four options including (1) respond briefly to literal and factual questions about the text, (2) identify main idea or discuss sequence of events in the story, (3) infer author's meaning by making connections within and between texts, and (4) build on each other's ideas citing evidence from text to support their assertions. We were interested in the latter two categories indicating a focus on inferences and using supporting evidence. This covariate was a product of time spent in text discussions and time in text discussion inferring or using supporting evidence.	.013–.078	.045	.018
Frequency of routine comprehension activities	This covariate is a factor of 3 survey responses to an item stem asking the following: "Below are a set of items pertaining to classroom text discussions. Considering constraints on your teaching please indicate the frequency students engage in each element in your text discussions." Items were answered on the following scale: never, rarely, sometimes, often, almost always. Three items formed a single factor explaining 57% of the variance in the items and included (1) students provide brief answers to comprehension questions, (2) students recount factual events in a story, and (3) students sequence events in a story.	3–5	3.95	.56

Table 2. (*Continued*)

Measures	Description	Range	Mean	*SD*
Logs:[b]				
Comprehension	Scale of comprehension items (activate prior knowledge, answer literal comprehension questions, examine story structure, analyze/synthesize, provide brief answers, students discuss text, provide extended answers, teacher-directed instruction, integrate writing into reading comprehension) across all logs contributed and weighted by time. This scale represents the exposure students had to comprehension across the year.	−1.73–2.21	0	1

[a] Data reduction was accomplished through factor analysis using SPSS 19.0 specifying principal axis factoring and an oblique rotation. A single factor was obtained explaining about 58% of the variance in the three covariates. Examination of the scree plot indicated a single factor was preferred.

[b] The frequency of days teachers reported using each of the 10 dichotomous items in comprehension was reported in Correnti et al. (2012). These dichotomous items were nested within days, and days were nested in teachers to construct scales from a measurement model (e.g., see Raudenbush et al., 1991). The models weighted each occasion by the number of minutes teachers reported teaching that day. From these measurement models we obtained an empirical Bayes residual from instructional scales estimated simultaneously including comprehension along with the aforementioned writing and integration of comprehension with writing scales.

ences from the rank order of classrooms based on students' average performance on the RTA. We sought evidence that performance on the RTA correlated with other known measures of student ability. Therefore, we compared and contrasted the scores for students on the two measures of student ability—the MSA and the RTA. Bivariate correlations between the average RTA score (mean score on the five dimensions) and the overall scale score on the MSA reading were .59 at the student level and .68 at the classroom level. Table 3 displays correlations among each dimension of the RTA with the MSA scale score; they range from .34 to .51 at the student level.

All of these correlations are statistically significant and connote a positive association between the MSA and RTA. Interestingly, rating categories based on the student response to the text (i.e., analysis and evidence) had lower correlations with the MSA than did more generic aspects of students' writing (i.e., organization, style, and MUGS). While all the associations were significant, the moderate correlations also suggest that the abilities measured by the two assessments do not overlap completely.

We explored the factor structure of the subscores of our two assessments. For the RTA the subscores included the scores for the five criteria on the rubric, and for the

Table 3. Bivariate Correlations[a] between RTA and MSA

	MSA Reading	Analysis	Evidence	Organization	Style	MUGS	RTA (Avg.)
MSA reading	–	.34	.41	.51	.51	.51	.59
Analysis	.32	–	.34	.51	.53	.41	.67
Evidence	.47	.56	–	.60	.50	.45	.73
Organization	.69	.57	.85	–	.63	.57	.83
Style	.72	.77	.67	.81	–	.68	.87
MUGS	.76	.60	.68	.82	.88	–	.80
RTA (avg.)	.68	.77	.86	.93	.94	.91	–

[a] Upper diagonal represents student-level correlations; lower diagonal represents classroom-level correlations.

MSA the subscores included the three scales previously defined. We ran a confirmatory factor analysis (CFA) using the software package MPlus6.12, assuming that the two tests measured different aspects of students' learning. Using all eight subscales, the model could not be identified because of the distribution of student scores on the MSA general subscale. We reran the analysis using all five subscales of the RTA and just two MSA subscales. This decision was made both because the general subscale had a poor distribution and because the general subscale featured multiple-choice items that did not directly relate to students' production of writing; in contrast, the two retained subscales contained multiple-choice comprehension items and brief constructed responses to literary and informational text passages, respectively. We compared the results obtained from both a one- and two-factor solution. The single-factor solution fit statistics ($\chi 2 = 138.42$, $df = 14$, $p < .000$; RMSEA = .144; CFI = .897; SPMR = .055) fail to demonstrate a good model fit even with modifications ($\chi 2 = 99.17$, $df = 12$, $p < .000$; RMSEA = .128; CFI = .928; SPMR = .047). The two-factor solution ($\chi 2 = 50.67$, $df = 13$, $p < .000$; RMSEA = .082; CFI = .969; SPMR = .028) demonstrates better model fit, which improves with modifications ($\chi 2 = 16.06$, $df = 11$, $p = .139$; RMSEA = .033; CFI = .996; SPMR = .017). Thus, even though both assessments purport to measure aspects of writing, our analyses demonstrate better model fit when each test is considered separately.

The results from the CFA carried several implications for our analyses. First, we were interested in exploring univariate hierarchical prediction models with the MSA and RTA considered separately. These allowed us to compare and contrast the RTA and MSA as outcomes in separate models. Here we were interested in comparing both the variance components as well as effects of our instructional covariates. Second, we were also interested in additional analyses, particularly given the fact that the RTA is unlikely to take precedence over the accountability test that districts are required to administer. We conducted two separate multivariate analyses in order to understand how both tests combined could provide additional information above and beyond the univariate analysis of the accountability test alone.

The first multilevel multivariate model built off of the results from the CFA. Here, we examined the five subscales of the RTA and the two writing subscales of the MSA (informational and literary, hereafter labeled $MSA_{inf+lit}$). The multilevel multivariate model allowed us to examine each as a separate test while accounting for the covariance between the two measures (Raudenbush, Rowan, & Kang, 1991; Snijders & Bosker, 1999; Thum, 1997). The psychometric phase of this model provided information about the correlation between the two tests after attenuating the scales for measurement error. Furthermore, these multivariate models allowed us to more accurately represent the reality of complex phenomena (Thum, 1997) by providing a means for comparing and contrasting effects of covariates across multiple outcomes (Hauck & Street, 2006; Hoffman & Rovine, 2007). For example, we examined students' opportunities for reasoning and extended writing as a predictor of each test. In particular we were interested to see if our measure of student reasoning and extended writing would allow us to demonstrate a teaching effect on the district's accountability test when examining just the two subscales incorporating elements of writing.

The second multilevel multivariate model examined an overall achievement score across the seven subscales of the RTA and $MSA_{inf+lit}$ while also examining the contrast between students' performance on the RTA relative to performance on the $MSA_{inf+lit}$. The purpose of this analysis was twofold. First, we wanted to examine a

general measure of writing and comprehension (including components of reading comprehension on the $MSA_{inf+lit}$ and listening comprehension on the RTA). Second, simultaneously, in the same model we wanted to examine whether our measure of students' opportunities for reasoning and extended writing was more potent for students' performance on one type of written assessment versus the other. In particular, we wanted to examine whether the RTA was more sensitive than the state accountability test to teaching that was theoretically aligned with skills needed to do well on the RTA.

Univariate Analyses Predicting Student Learning

Using HLM7.0, we first examined each outcome separately in a two-level covariate adjusted model. The general form of this model is:

$$\text{Level 1 (students): (achieve)}_{ij} = \pi_{oj} + \pi_{pj} * (A_{pi}) + e_{oj}, \quad (1.1)$$

$$\text{Level 2 (teachers): } \pi_{oj} = \beta_{po} + \sum_{q=1}^{Q} \beta_{pq} X_{qj} + r_{oj}, \quad (1.2)$$

where $(\text{achieve})_{ij}$ is the achievement of student i in teacher j's classroom; π_{oj} is the average achievement across all students; (A_{pi}) is a set of (p) student-level covariates for student i; π_{pj} is the effect of each (p) student-level covariate on achievement; β_{po} is the average achievement across all classrooms; X_{qj} is a set of teacher and classroom covariates; β_{pq} is the effect of teacher and classroom covariates on achievement, and where e_{oj} and r_{oj} are independent normal residual errors.

Instructional covariates in univariate analyses. We examined whether covariates derived from the instructional data have logical relationships with different measures of student learning. Since much prior research has demonstrated the importance of curricular alignment with the assessment (D'Agostino et al., 2007; Ruiz-Primo et al., 2002), we attempted to test for the presence of alignment between measures of teaching (the enacted curriculum) and measures of learning. Here we examined students' opportunities for comprehension and students' opportunities for reasoning and extended writing as covariates predicting student learning on the MSA and RTA. Items on the MSA are a mix of multiple-choice comprehension questions, brief constructed responses (in response to a short passage), and other items such as vocabulary; hence we expected that students with greater opportunities for comprehension would score higher on the MSA because we thought the total scale score reflected general comprehension skills. In contrast, we expected that students with greater opportunities for reasoning and extended writing would perform better on the RTA than students with fewer opportunities to develop these skills. To test these hypotheses, we compared and contrasted results from univariate models on the MSA and RTA, respectively.

Multivariate Multilevel Analyses Predicting Student Learning

One reason for examining the multivariate models is that we know student performance is related on our two tests—the state test (MSA) and the response-to-text assessment (RTA). We examined a three-level hierarchical linear model using HLM 7.0 (Raudenbush & Bryk, 2002). At level 1, this is a measurement model that describes

the subscores contributing to each achievement scale and examines the measurement error variation in the true-score estimation of the achievement scales. Levels 2 (student level) and 3 (classroom level) of this analysis then are essentially a multivariate two-level model for the latent scale scores of achievement. These models are further described in Appendix B. Analyses of the variance components and psychometric data from the measurement model help evaluate whether the scales created from RTA and MSA subscores reliably distinguish between students and classrooms on each scale.

Instructional covariates in multivariate analyses. To follow-up on findings of our univariate analyses, we first included opportunities for reasoning and extended writing in a multilevel multivariate model as a classroom level predictor of the RTA $(\beta_{105})_1$ and $\text{MSA}_{\text{inf+lit}}$ $(\beta_{205})_1$. In a second multivariate multilevel model, we examined whether the same measure of instruction predicted overall achievement on the RTA and $\text{MSA}_{\text{inf+lit}}$ together $(\beta_{105})_2$ and the effect of students' opportunities for reasoning and extended writing on the contrast between performance on the RTA versus $\text{MSA}_{\text{inf+lit}}$ $(\beta_{205})_2$.

Results

We examined a series of prediction models to investigate our initial claims. These claims were essentially twofold. First, the RTA will distinguish among classrooms in students' ability to write analytically in response to a text. Second, performance on the RTA is aligned in logical ways with measures of the enacted curriculum, specifically measures of students' opportunities to both think critically about texts in discussions and in class activities, and to produce extended writing. The first claim is partially addressed by understanding whether there is sufficient variance between classrooms to detect relationships. At the same time, both of the claims utilize information from prediction models to gauge the extent to which we can infer that the observed variance is indicative of students' analytic writing skills.

Student Performance on the RTA

We begin by describing students' performance on the RTA to provide an overall picture of what the ratings reveal about students' response-to-text writing skills, as well as to provide information about the ratings that could be useful for interpreting our quantitative results. As shown in Figure 1, the RTA ratings showed substantial variation in student responses across the five dimensions (i.e., criterion). The modal response for each dimension was 2, indicating room for improvement across students in general in the quality of their responses; however, means and distributions for the individual rating scales varied considerably. Analysis was the most highly skewed of the five scales and had the only mean below 2 ($M = 1.88$). Only 21% of the students scored a 3 or a 4, indicating that their responses showed evidence that they understood the purpose of the text and could synthesize ideas in the text. These students made a clear inference and articulated it in the form of a valid and insightful claim in direct response to the prompt. The majority of students, in contrast, demonstrated a very limited understanding of the text and had a great deal of difficulty inferring meaning from the text. Students may, for example, merely summarize the text or restate an obvious or given conclusion from the selection. Such responses may

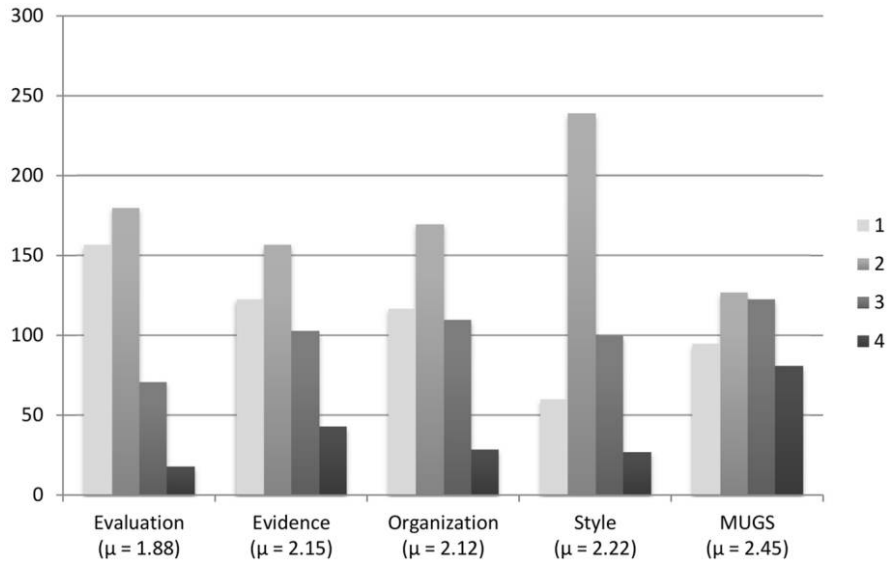

Figure 1. Distribution of student scores on five scoring criteria of the RTA rubric.

also include partial or unclear claims that are not explicitly articulated. Evidence, Organization, and Style all had a slight skew toward the lower end of the distribution and a similar mean; otherwise, they were fairly normally distributed. Students that scored 3 or 4 on Evidence provided a sufficient number of pieces of evidence and the evidence selected was of concrete, specific events or utterances in the text. Furthermore, the evidence was relevant and significant to the claim being argued. The typical response, with a score of 2, however, indicated that students used limited pieces of evidence that were often unclearly related to the argument or only referred generally to parts of the text. While our scale rewarded students for constructing multiple paragraphs, having clear topic and concluding sentences, and demonstrating logical flow between sentences and paragraphs, we found that the majority of students gave a single-paragraph response. Students tended to provide some version of a topic sentence that attempted to address the prompt; however, the writing tended to end abruptly, without a conclusion. The body of the paragraph may be underdeveloped (some responses consisted of only one or two sentences total) or contain ideas that, while on topic, are not organized coherently. Finally, students typically used expected word choices, simple sentence structures, and basic connectives instead of more advanced academic language. A typical response also featured persistent grammatical and mechanical errors, such as run-on sentences or fragments.

Annotated sample student responses are included in Appendix C. In the interest of brevity, in order to address both prompts and to show each scale value of the rubric (1–4), we have provided two pieces of writing for each prompt. Furthermore, in the interest of space, we have selected responses that earned the same score on all five criteria; this is a rare phenomenon in our sample. The four pieces provide a sampling of the quality of student writing to help interpret and understand our quantitative findings. They demonstrate some obvious qualitative differences across students' reasoning and writing. For one, the sample responses demonstrate some of the ways students' strategies for answering our prompt went awry. For example, most students had a heavy reliance on the text itself and thus had trouble generalizing from

the text to take an analytic stance. The samples also demonstrate what it looked like when students used incomplete evidence (which was also a frequent occurrence), as in the focus in sample C on the before condition of Sauri to the exclusion of the after condition. Finally, in addition to demonstrating the variance in students' performances, the annotated pieces of student writing provide the reader with a sense of how our raters applied the rubrics provided in Appendix A.

Comparing and Contrasting Univariate Hierarchical Models on the MSA and RTA

Variance components analysis. Tables 4 and 5 display the results for the random effects for each of the univariate models. We focus our immediate attention on the null model containing no covariates and the model with only background characteristics of students and teachers (located in the first two columns of each table). For both outcomes, background characteristics explain a large proportion of the variance at both the student (51% reduction in variance for the MSA and 32% for the RTA[8]) and teacher (68% reduction in variance for the MSA and 38% for the RTA) levels. It is important to note one difference: the background characteristics explained more variance in student performance on the MSA at both levels than the background characteristics did on the RTA. Even for the MSA, however, the chi-square statistics of variance remaining between classrooms reveal significant differences in classroom performance after adjusting for student and teacher characteris-

Table 4. Univariate Hierarchical Linear Models of Maryland School Assessment (MSA)

	Coefficient (*SE*)			
	Null Model	Background Characteristics	Comprehension Factor	Extended Writing Factor
Teacher-level fixed effects:				
Intercept	−.04 (.13)	−.01 (.08)	−.02 (.08)	−.01 (.08)
Grade		.25 (.14)	.24 (.14)	.23 (.14)
Ph.D.		.25 (.46)	.18 (.55)	.24 (.49)
Advanced certification		−.21 (.26)	−.24 (.28)	−.22 (.27)
Years experience		−.01 (.01)	−.01 (.01)	−.01 (.01)
Instruction			.05 (.12)	.08 (.10)
Student-level fixed effects:				
Absences		−.01** (.01)	−.01** (.01)	−.01** (.01)
Hispanic		.20 (.17)	.19 (.17)	.19 (.17)
Black		−.04 (.19)	−.03 (.19)	−.03 (.19)
Native American		−.19 (.23)	−.18 (.23)	−.18 (.23)
Asian		.22 (.23)	.22 (.23)	.22 (.23)
Free lunch		−.10 (.08)	−.10 (.08)	−.10 (.08)
Reduced-price lunch		−.09 (.10)	−.09 (.10)	−.09 (.10)
Individualized education plan		.14 (.11)	.14 (.11)	.14 (.11)
Prior reading achievement		.56*** (.05)	.56*** (.05)	.56*** (.05)
Prior math achievement		.23***(.05)	.23*** (.05)	.23*** (.05)
Random effects:				
Between-classroom variance (τ_{β_0})	.28	.09	.09	.09
χ^2 (*df*)	175.76 (17)	82.48 (13)	79.17 (12)	77.02 (12)
Between-student variance (σ^2)	.72	.35	.35	.35

** $p < .01$.
*** $p < .001$.

Table 5. Univariate Hierarchical Linear Models of Response-to-Text Assessments (RTA)

	Coefficient (*SE*)			
	Null Model	Background Characteristics	Comprehension Factor	Extended Writing Factor
Teacher-level fixed effects:				
Intercept	.00 (.13)	.05 (.11)	.05 (.11)	.05 (.09)
Grade		−.05 (.19)	−.07 (.18)	−.13 (.16)
Ph.D.		.03 (.55)	−.14 (.63)	.02 (.50)
Advanced certification		.35 (.29)	.29 (.30)	.33 (.24)
Years experience		.00 (.02)	.00 (.02)	−.01 (.01)
Instruction			.18 (.14)	.30* (.11)
Student-level fixed effects:				
Absences		.00 (.01)	.00 (.01)	.00 (.01)
Hispanic		.09 (.20)	.09 (.20)	.08 (.20)
Black		.13 (.22)	.13 (.22)	.14 (.22)
Native American		−.08 (.26)	−.07 (.26)	−.08 (.26)
Asian		.35 (.26)	.36 (.26)	.35 (.26)
Free lunch		−.17* (.09)	−.16* (.09)	−.17* (.09)
Reduced-price lunch		−.18 (.12)	−.17 (.12)	−.17 (.12)
Individualized educational plan		−.28* (.12)	−.28* (.12)	−.29* (.12)
Prior reading achievement		.33*** (.06)	.33*** (.06)	.32*** (.05)
Prior math achievement		.27*** (.06)	.27*** (.06)	.26*** (.06)
Random effects:				
Between-classroom variance ($\tau_{\beta o}$)	.31	.19	.18	.11
χ^2 (*df*)	217.66 (17)	123.59 (13)	110.16 (12)	68.85 (12)
Between-student variance (σ^2)	.68	.46	.46	.46

* $p < .05$.
*** $p < .001$.

tics ($\chi^2 = 82.43$, $df = 13$, $p < .001$). Likewise, the RTA also revealed significant differences in variance remaining between classrooms ($\chi^2 = 123.59$, $df = 13$, $p < .001$) after adjusting for background. Finally, while both outcomes had significant between-classroom variance after adjusting for background, the proportion of variance between classrooms was lower for the MSA (ICC = .21) than it was for the RTA (ICC = .29).[9]

Effects of covariates. Some covariates in Tables 4 and 5 demonstrate consistent effects across the MSA and RTA. Background characteristics, for example, often had similar effects in both models. Prior achievement in both reading and math were highly significant on both outcomes. It bears noting that the magnitude of the effect for prior math achievement was about the same as the magnitude for prior reading achievement on the RTA, which was not the case for the MSA, where reading had a larger effect. Free lunch seems to have a fairly consistent effect across the models, but is statistically significant on the RTA ($p = .041$) but not on the MSA ($p = .190$).

A few differences also emerge for the covariates. Grade predicts higher scores on the MSA, which is logical since the MSA is equated across grades, whereas no grade differences were seen on the RTA where it was not possible to equate scores across grades. Second, students with an IEP scored similarly to their peers on the MSA but had lower performance on the RTA. No test accommodations were provided on the RTA, whereas test accommodations were available for the MSA. Finally, student absences predicted lower performance on the MSA but not the RTA. On balance, the pattern of results demonstrated that the RTA was sensibly related to student back-

ground characteristics that we would expect it to have an association with (e.g., prior achievement, free lunch status, and special needs) but not to others we would not expect it to have an association with (e.g., grade). Finally, no teacher-level background characteristics that were included as proxies for knowledge and experience were significant for either outcome.

Instructional covariates. The pattern of results for the instructional covariates provides another point of comparison for the performance of students on each test. In the right-hand column of Tables 4 and 5 we display the effects of students' opportunities for comprehension and for student reasoning and extended writing, each included in separate models.

The RTA was sensitive to instructional effects in the ways we predicted, but neither the factor measuring students' opportunities for comprehension nor for reasoning and extended writing was a significant predictor of student performance on the MSA. Moreover, in the MSA models the addition of either instructional covariate resulted in no further reduction in between-classroom variance beyond student and classroom background factors. Thus, while the coefficients of all instructional effects were positive for both outcomes, only the RTA was sensitive to instructional effects, and only then to one of the two instructional factors—students' opportunities for reasoning and extended writing ($ES = .40$).

Instructional factors explained a portion of the variance remaining between classrooms on the RTA after adjusting for student background. Students' opportunities for comprehension, despite being nonsignificant, explained about 6% of the remaining between-classroom variance after student background characteristics were accounted for.[10] Subsequently, the model examining students' opportunities for reasoning and extended writing as a covariate explained about 42% of the remaining between-classroom variance.

Comparing and Contrasting Results from Multilevel Multivariate Hierarchical Models

Psychometric findings. In the multilevel multivariate hierarchical linear model (HLM) we examined the relationship between $MSA_{inf+lit}$ (the combination of the MSA subscales on the informational and literary subsections containing both comprehension multiple-choice items and students' brief constructed responses) and the RTA. This model allows for a variance decomposition, treating σ^2 as measurement error and decomposing the remaining "true score variance" into portions remaining between students within classrooms and between classrooms. These findings illustrate the potential of the assessment to reliably distinguish between students on the $MSA_{inf+lit}$ (.62) and RTA (.78), respectively, and also between classrooms on the $MSA_{inf+lit}$ (.87) and RTA (.90). These reliability estimates provide a measure of the internal consistency of the scales relative to the measurement error. In each case, variance exists to be explained by differences between both individuals and classrooms. Correlations were also obtained between the RTA and $MSA_{inf+lit}$ via this method. "True score" correlations between the ability scales at the student ($r = .76$) and teacher ($r = .73$) level were slightly higher in the measurement model than they were in bivariate correlations reported in Table 3 (.59 and .68), which is natural given that this model treats deviations from the average as residual measurement error.

Effects of opportunities for reasoning and extended writing. Findings for multivariate model 1 in Table 6 resemble the findings from univariate analyses. Students' opportunities for reasoning and extended writing were predictive of performance on the RTA after adjusting for student and classroom covariates ($\beta_{105} = .244$; $ES = .46$) but not significantly predictive of the $MSA_{inf+lit}$ ($p = .681$). One difference in the models is that 77% of the variance in $MSA_{inf+lit}$ is accounted for by prior achievement on the MSA and student background, whereas student background accounts for about 45% of the variance in RTA scores.[11] Two additional findings merit brief consideration. First, no other teacher or classroom covariates were associated with achievement on either measure. Second, the addition of our measure of opportunities for reasoning and extended writing explained about half of the variance in the RTA (50%) when compared to the model adjusting for student background characteristics alone.

Findings from multivariate model 2 confirm the notion that a focus on extended writing was associated with higher performance on combined achievement on the $MSA_{inf+lit}$ and RTA. Students' opportunities for reasoning and extended writing were associated with higher overall achievement scores ($\beta_{105} = .164$; $ES = .38$). Additionally, opportunities for reasoning and extended writing were also predictive of the contrast in performance on the RTA relative to the $MSA_{inf+lit}$. Students with greater opportunities for reasoning and extended writing have relatively higher performance on the RTA versus the $MSA_{inf+lit}$ ($\beta_{205} = .401$; $ES = .46$). Combined with findings from multivariate model 1, we found that greater opportunities for reasoning and extended writing were associated with better performance overall, but in particular with higher scores on the RTA relative to performance on the $MSA_{inf+lit}$.

Discussion

The purpose of our study was to conduct a validity investigation of an assessment of students' analytic text-based writing skills. The motivation for creating this assessment (the RTA) was twofold. First, the majority of current assessments measure reading comprehension and writing as separate skills, which produces satisfying psychometrics but is not well aligned with the intent of the CCSS. Thus, in our view current assessments remain only modestly aligned with the particular writing skills that students need to succeed at higher levels of schooling. Second, we reasoned that our assessment would be more sensitive to teaching that emphasizes text analysis and extended writing (i.e., higher-level literacy skills) than most readily available assessments. As such, our assessment would be useful in research focused on high-level teaching-learning connections, as well as research investigating the effect of educational reforms on teaching and students' literacy skills. Overall, our findings provide supporting evidence that, for a small sample of classrooms at least, the RTA can simultaneously measure students' performance on a task that combines text analysis, use of evidence, and extended writing. Moreover, we were able to measure teaching practice that was theoretically aligned with the development of students' analytical text-based writing skill and such instruction was also associated with classroom performance on the RTA.

Table 6. Tests of Association between the RTA, $MSA_{inf+lit}$, and Opportunities for Reasoning and Extended Writing

	Multivariate Model 1				Multivariate Model 2			
	RTA $(\beta_{105})_1$		$MSA_{inf+lit}$ $(\beta_{205})_1$		RTA + $MSA_{inf+lit}$ $(\beta_{105})_2$		Contrast RTA/$MSA_{inf+lit}$ $(\beta_{205})_2$	
	Coeff.	*SE*	Coeff.	*SE*	Coeff.	*SE*	Coeff.	*SE*
Teacher-level fixed effects:								
Intercept	.024	.056	−.031	.056	.004	.051	.104	.092
Grade	−.094	.107	.162	.102	.002	.096	−.481**	.162
Ph.D.	.080	.344	.152	.367	.107	.337	−.135	.403
Advanced professional certification	.261	.159	−.141	.206	.110	.159	.753*	.308
Years experience	−.008	.010	−.005	.009	−.007	.009	−.006	.014
OTL reason/write	.245***	.069	.030	.072	.164*	.064	.401**	.116
Student-level fixed effects:								
Hispanic	.023	.140	.289	.157	.123	.116	−.497	.346
Black	.116	.162	.044	.175	.089	.135	.135	.381
Native American	−.036	.182	−.182	.206	−.091	.154	.273	.441
Asian	.213	.193	.215	.208	.214	.160	−.003	.459
Free lunch	−.127*	.062	−.084	.069	−.111*	.053	−.080	.148
Reduced lunch	−.030	.091	.123	.102	.028	.077	−.287	.217
Individualized educational plan	−.227*	.088	.165	.102	−.080	.074	−.733***	.221
Reading prior	.256***	.040	.476***	.046	.339***	.034	−.411***	.100
Math prior	.190***	.043	.217***	.048	.200***	.036	−.050	.104
Absences	−.004	.004	−.015**	.005	−.008*	.003	.021*	.010
Random effects:	Null	Pred.	Null	Pred.	Null	Pred.	Null	Pred.
Between-classroom variance	$r_{10} = .176$	$r_{10} = .044$	$r_{20} = .186$	$r_{20} = .041$	$r_{10} = .156$	$r_{10} = .038$	$r_{20} = .350$	$r_{20} = .079$
Between-student variance	$e_1 = .324$	$e_1 = .179$	$e_2 = .382$	$e_2 = .186$	$e_1 = .305$	$e_1 = .117$	$e_2 = .614$	$e_2 = .427$
Between-item variance (σ^2)	.47				.47			

* $p < .05$.
** $p < .01$.
*** $p < .001$.

Qualities of Student Scores on the RTA

A key finding from our study is that students' analytic text-based writing skills were weak overall. Relatively few students were able to generate a clear inference from the text, or use appropriate evidence from the text to support their claims. Students' writing also showed weaknesses with respect to their use of the vocabulary and more sophisticated sentence structures that characterize an academic voice (Crosson, Matsumura, Correnti, & Arlotta-Guerrero, 2012). Results are not surprising given the generally poor state of writing instruction in our country, as well as research showing that relatively few students are prepared to do the kinds of writing that are required for success at higher levels of schooling and in the workplace (American Diploma Project, 2004; Applebee & Langer, 2009). Our results also suggest that students will likely have a great deal of difficulty meeting the writing standards set out in the CCSS that emphasize formulating and supporting text-based arguments. Substantive changes in classroom practice, and professional learning opportunities for teachers, will be needed.

For each of our five criteria (analysis, evidence, organization, style, and MUGS) we demonstrated moderate to high agreement rates based on Cohen's kappa between two trained raters. Notably, these rates were highest for the more objective categories (rating aspects of students' writing quality) and lower for more subjective categories (rating students' analysis of text and use of supporting evidence). Interestingly, we found that students rarely scored the same on all five of our rubric criteria.[12] Correlations among the dimensions indicated that scores on the dimensions varied within students, suggesting that multiple criteria were necessary for capturing differences among students in analytic writing skills. Understanding how and why students differ in their analytic text-based writing, as well as rater reliability among dimensions, are important topics for further study.

Supporting the first claim of our validity investigation, our results provide evidence that the RTA ratings were sensitive to variation between both students and classrooms. We compared and contrasted student scores on the RTA and MSA for our sample of 426 students nested in 18 classrooms. Both tests demonstrated reliable between-student and between-classroom variance in our fully unconditional and conditional HLM models. Furthermore, the test scores were correlated at the student and classroom levels, although the strength of association was not so great that the tests seemed to duplicate one another. The moderate correlation is not surprising given that the MSA does ask students to write brief constructed responses and therefore measures student writing in addition to other skills such as reading comprehension, vocabulary, and word study. Indeed, correlations between the overall MSA scale score and each dimension of the RTA demonstrate that the correlation between MSA and RTA dimensions was highest where more generic aspects of student writing were rated (i.e., organization, style, and MUGS) and lower where reasoning and use of supporting evidence were rated (i.e., analysis and evidence).

One possible explanation for this finding is that the dimensions of the RTA with less overlap with the MSA capture aspects of student learning that are difficult to evaluate and not often included in formal assessments. If this were the case, outcomes on the RTA could serve to inform both research and policy. Another possible explanation is that the low correlations are due to greater measurement error resulting from greater subjectivity of the rating judgments required for those dimensions

or to other sources of error such as task sampling. Replication studies are necessary to understand whether the correlations demonstrated here are reproduced in other samples. If these correlations are replicated, it will be necessary to uncover what the most likely explanations are for them.

In the meantime, we remain cautiously optimistic that the RTA serves an important function in research on student learning. Take, for example, the finding that the RTA demonstrated a higher proportion of between-classroom variance in student learning than the MSA after adjusting for student background characteristics, including prior reading and mathematics achievement on the MSA. One possible explanation is that the RTA presents an alternative (or additional) assessment to the accountability test used in the district (the MSA). It is thus more apt to be sensitive to natural variation between classrooms in students' ability because, whereas classrooms prepare for the MSA, they do not prepare for the RTA.

Another possible explanation is that the lack of a prior measure of student writing and the use of the prior MSA as the only adjustment for student background explain proportionally more of the variance in the MSA model versus the RTA model. However, if the tests were duplicating one another, then it is unlikely that either test would demonstrate greater between-classroom variance, and the effects of prior MSA scores on both outcomes would be similar. Further, while individually there are potential alternative explanations for all of the above findings, taken together, findings about the relationship between students' RTA criterion scores with the MSA suggest that the RTA deserves careful consideration when researchers are seeking to understand students' abilities to reason about the content in text(s) and express their ideas in writing.

Findings of Instructional Covariates on Classroom Performance

Analyses examining covariates related to classroom performance on the MSA and RTA demonstrate that the same instructional measures had different patterns of relationships with the two student outcomes. The major finding was that classroom performance on the RTA was related to a measure of instruction that was theoretically aligned in both content and format to the RTA. Despite low power to detect effects, we observed a significant association between our measure of teaching reasoning and extended writing, and the RTA ($ES = .46$). The MSA, meanwhile, was not sensitive to students' opportunities for reasoning and extended writing. In addition, the RTA was also not sensitive to the reading comprehension covariate. These intuitive findings should be considered in light of research showing the difficulty of identifying objective instructional actions associated with student outcomes (Carlisle, Kelcey, Beribitsky, & Phelps, 2011; Rosenshine & Furst, 1973).

The association we found between instruction that emphasizes text analysis and extended writing and the RTA suggests that response-to-text writing is a domain of instruction and learning fruitful for further exploration. Researchers should continue both to examine the nature of this association (i.e., the extent to which one causes the other) and also to understand the sources of variation for teaching student reasoning and extended writing. This is especially timely because the teaching-learning connection is currently omitted from accountability frameworks, and because reasoning and extended writing are skills to be incorporated in the assessments that are being developed to align with the Common Core State Standards. The im-

portance of generating analytic writing in response to text is likely to gain in importance if the CCSS remain in place. Students' abilities to reason and write extended responses are essential skills for college and career readiness.

More generally, the research findings further demonstrate the importance of attending to principles of specificity and overlap when exploring teaching-learning associations. The findings suggest that alternative assessments have the potential to reveal relationships between teaching and learning that are not obtainable through current accountability assessments. Thus, the sampled domain of student reasoning/extended writing is important because other alternative assessments sampling from a target domain different from the accountability test may be able to demonstrate further unique teaching-learning associations. Therefore, it is important to be deliberate in our choice of student learning outcomes as these will condition the teaching effects that are available for researchers to uncover.

A final contribution of this work is in demonstrating the importance of simply having multiple student learning outcomes. We have demonstrated one way multiple student outcomes can be used in both univariate and multivariate prediction models. In multivariate models we demonstrated how our measure of students' opportunities for reasoning and extended writing predicted a general outcome of student learning combining portions of the MSA ($MSA_{inf+lit}$) with the RTA, and how it also predicted the contrast in classroom performance on the RTA relative to classroom performance on the $MSA_{inf+lit}$. These analyses demonstrate how researchers and districts could use assessments in addition to those required for accountability purposes to learn how natural variation in measures of teaching are associated with better classroom performance. In turn, such knowledge could be useful to researchers, individual practitioners, and interventionists thinking about school improvement.

Limitations and Directions for Future Research

Our approach to assessment on the RTA was to measure students' ability to express in writing how they reasoned about a text and provided evidence to support their assertions. As such, it included no explicit measurement of students' reading fluency or their reading comprehension. We made intentional choices (e.g., having the teacher read the text aloud and ask questions) to emulate authentic classroom instruction and to hone in on the constructs we were most interested in measuring—the combination of reasoning and writing. Other researchers may choose to have students read the text themselves, or perhaps choose not to provide conversational supports for literal comprehension of the text and thus sample more completely from the target domain to assess not only reasoning and writing but also reading fluency and reading comprehension.

While we have noted the differences between the RTA and MSA in our research findings, further study of differences between the RTA and other state standardized assessments is warranted. One limitation of the work presented here is the difference in test administrations and, in particular, the timing of each assessment. Thus, while the RTA seems more sensitive to our instructional measures, a contributing factor may have been that the RTA was administered 2 months after the MSA. Thus, students could demonstrate an additional 2 months of accumulated learning opportunities on the RTA. Researchers should seek to understand how such timing could

affect the extent to which standardized tests are sensitive to measures of teaching because it has implications for research findings when exploring teaching-learning associations. It also has potential implications for practitioners as such tests are currently used as a component of teaching evaluation systems being put in place in many schools (Baker et al., 2010).

Conclusion

We chose to examine students' ability to reason about a text and then create an extended analytic written response because we see this as an important skill that should merit greater emphasis in the language arts curricula, even as early as grade 4. In our research we identified an association between teaching reasoning and extended writing and our response-to-text assessment. We think this is an area ripe for future study and, furthermore, our findings suggest the importance of identifying student learning outcome(s) aligned with specific elements of "effective teaching." Through our analytic methods we demonstrated one method of examining both univariate and multivariate prediction models, which highlighted the utility of having multiple student learning outcomes. Careful consideration of student outcomes is critical going forward because the process of identifying teaching-learning associations forms the basis for building a teaching knowledge base that can guide professional learning and inform both policies and instructional interventions.

Appendix A

Table A1. RTA Scoring Rubric (with Decision Rules)

	Analysis	Evidence	Organization	Style	Mechanics/Usage/Grammar/Spelling
4	• Demonstrates a clear understanding of the (purpose of the) literary work • Exhibits original insight and draws meaningful conclusions or demonstrates sophisticated and succinct synthesis of ideas • Inference/insight is clearly, explicitly articulated • Inference is elaborated upon, not just stated in a sentence or a few words in the beginning • Addresses the prompt	• Selects detailed, precise, and significant evidence from the text • Demonstrates integral use of selected details from the text to support and extend key idea	• Focuses on the main idea • Demonstrates logical and seamless flow from sentence to sentence and idea to idea • Has a strong sense of beginning, middle, and end • Beginning and end must relate closely to same key idea • Must feature multiple appropriate paragraphs	• Features varied sentence lengths and structures, including complex structures • Uses tier 2 vocabulary multiple times • Uses sophisticated connectives[a] multiple times correctly • Features a number of sophisticated or original phrases	• Features errors that do not detract from communication of ideas • Features very few minor errors or a few "sophisticated" errors
3	• Demonstrates adequate understanding of the (purpose of the) literary work • Makes multiple valid and insightful inferences or demonstrates synthesis of ideas • Inference/insight is clearly, explicitly articulated, not requiring interpretation by reader • Addresses the prompt	• Selects appropriate and concrete, specific evidence from the text • Demonstrates use of selected details from the text to support key idea	• Adheres to the main idea • Demonstrates logical flow from sentence to sentence and idea to idea • Has a clear sense of beginning, middle, and end • Makes a clear attempt at presenting a key idea in beginning and end, but may be weak in that they do not relate closely, and may be missing a beginning or end, but organization is strong otherwise (e.g., multiple paragraphs, well-developed middle, etc.)	• Features varied sentence structure • Uses tier 2 vocabulary a few times • Uses sophisticated connectives a few times • Features some sophisticated or original phrases	• Features errors that detract somewhat from communication of ideas • Features occasional minor errors

Continued

Table A1. (*Continued*)

	Analysis	Evidence	Organization	Style	Mechanics/Usage/Grammar/Spelling
2	• Demonstrates a limited understanding of the (purpose of the) literary work • Makes credible but obvious and basic inferences, which may contain minor errors • Inference often based on one sentence or even just a few words • Addresses the prompt	• Selects some appropriate but general evidence from the text; may contain a factual error or omission • Evidence provided may be list-like in a sentence or so, not expanded upon • Demonstrates limited development or use of selected evidence	• Has serious organizational problems • Has some uneven flow or some unclear passages • Has limited sense of beginning, middle, and end • Missing a beginning or end, and middle exists but may be short on development	• Uses mechanical, almost robotic (e.g., "chaining") sentence constructions • Features dull and repetitive word choice • Uses simple connectives correctly	• Features errors that interfere with communication of ideas/impede clear understanding of meaning • Features several major or simplistic errors; number of errors may be disproportionate to amount of text
1	• Demonstrates little understanding of the (purpose of the) literary work • Makes many invalid, irrelevant, inaccurate, or inadequate inferences • Substitutes summary for inference or fails to address the prompt altogether	• Selects inappropriate or little evidence from the text; may have serious factual errors and omissions • Demonstrates little or no development or use of selected evidence • May consist mainly of sentences copied from the text or may summarize entire text instead of locating evidence	• Has little or no order • May feature a rambling collection of thoughts or list-like ideas with little or no flow • Has little or no sense of beginning, middle, and end	• Features incomplete sentences; may only consist of a string of words conveying little or no information • Contains confusing or inaccurate word choice • Uses simple connectives incorrectly or no connectives • Features minimal (original) writing (1–2 sentences), too little to adequately assess	• Features errors that seriously interfere with communication of ideas/impede understanding of meaning • Frequently features major and simplistic errors; number of errors may be disproportionate to amount of text • Features minimal (original) writing (1–2 sentences), too little to adequately assess

[a] Additive (e.g., "additionally" and "also"), causal ("because" and "since"), adversative ("although," "nonetheless"), temporal (first, second or first, then, finally).

Appendix B

Multivariate Multilevel Model 1

Before running the models, student subscores were standardized to have a mean of zero and standard deviation of one. In addition, we first ran a null model to examine whether the scale variances were equivalent so that the writing and MSA scales could be easily contrasted (see, e.g., Raudenbush, Rowan, & Kang, 1991). The psychometric phase of these models also allowed us to calculate between-student and between-teacher reliabilities in addition to examining which covariates predict higher scores when considering each scale simultaneously as independent outcomes. The level 1 model is described below:

$$\text{achieve}_{mij} = \psi_{1ij} * (\text{RTA}_{mij}) + \psi_{2ij} * ((\text{MSA}_{\text{inf+lit}})_{mij}) + \varepsilon_{mij}, \quad (2.1)$$

where achieve_{mij} is the achievement subscore for scale m for student i in classroom j; RTA_{mij} is a dummy indicator demarcating the five subscores (or scoring criteria) of the writing rubric; ψ_{1ij} is the average writing achievement for student i in classroom j; $(\text{MSA}_{\text{inf+lit}})_{mij}$ is a dummy indicator demarcating the informational and literary subscales of the MSA; ψ_{2ij} is the average $\text{MSA}_{\text{inf+lit}}$ achievement for student i in classroom j; and ε_{mij} is the measurement error for dimension m for student i in classroom j. The level 2 (student-level) model is written as follows:

$$\begin{aligned} \psi_{1ij} &= \pi_{10j} + \pi_{1pj} * (A_{pi}) + e_{1ij} \quad (2.2) \\ \psi_{2ij} &= \pi_{20j} + \pi_{2pj} * (A_{pi}) + e_{2ij}, \end{aligned}$$

where π_{10j} is the average RTA achievement for students in classroom j; A_{pi} is a set of (p) covariates for student i; π_{1pj} is the effect of student-level covariates on RTA achievement; e_{1ij} is residual error normally distributed with mean of 0 and standard deviation of unity; π_{20j} is the average $\text{MSA}_{\text{inf+lit}}$ achievement for students in classroom j; A_{pi} is a set of (p) covariates for student i; π_{2pj} is the effect of student-level covariates on $\text{MSA}_{\text{inf+lit}}$ achievement; and e_{2ij} is residual error normally distributed with mean of 0 and standard deviation of unity. The level 3 (teacher-level) model is written as

$$\pi_{10j} = \beta_{100} + \sum_{q=1}^{4} \beta_{1pq} \quad X_q + \beta_{105}(\text{OTL for reasoning and writing}_j) + r_{10j} \quad (2.3)$$

$$\pi_{20j} = \beta_{200} + \sum_{q=1}^{4} \beta_{2pq} \quad X_q + \beta_{205}(\text{OTL for reasoning and writing}_j) + r_{20j},$$

where β_{100} is the average RTA achievement across all classrooms; X_q is a set of (q) teacher and classroom characteristics; β_{1pq} is the effect of teacher and classroom characteristics on RTA achievement; β_{105} is the effect of students' opportunities for reasoning and writing on the RTA; r_{10j} is residual error normally distributed with a mean of 0 and a standard deviation of unity; β_{200} is the average MSA achievement across all classrooms; X_q is a set of (q) teacher and classroom characteristics; β_{2pq} is the effect of teacher and classroom characteristics on RTA achievement; β_{205} is the effect of students' opportunities for reasoning and writing on the $\text{MSA}_{\text{inf+lit}}$; and r_{20j} is residual error normally distributed with a mean of 0 and a standard deviation of unity.

Multivariate Multilevel Model 2

A similar analysis examined a second multivariate model. This model (see eqq. 3.1 through 3.3 below) takes a similar form to the previous multilevel multivariate model with the exception that the scales no longer represent each test separately.

$$\text{achieve}_{mij} = \psi_{1ij} * ((\text{MSA}_{\text{inf}+\text{lit}} + \text{RTA})_{mij}) + \psi_{2ij} * (\text{contrast}_{mij}) + \varepsilon_{mij}, \quad (3.1)$$

$$\psi_{1ij} = \pi_{10j} + \pi_{1pj} * (A_{pi}) + e_{1ij} \quad (3.2)$$
$$\psi_{2ij} = \pi_{20j} + \pi_{2pj} * (A_{pi}) + e_{2ij},$$

$$\pi_{10j} = \beta_{100} + \sum_{q=1}^{4} \beta_{1pq} X_q + \beta_{105}(\text{OTL for reasoning and writing}_j) + r_{10j}, \quad (3.3)$$

$$\pi_{20j} = \beta_{200} + \sum_{q=1}^{4} \beta_{2pq} X_q + \beta_{205}(\text{OTL for reasoning and writing}_j) + r_{20j}.$$

Instead, the first scale considers achievement on the $\text{MSA}_{\text{inf+lit}}$ and RTA together ($(\text{MSA}_{\text{inf+lit}} + \text{RTA})_{mij}$). All seven achievement subscores of achievement are dummy coded 1 for this scale and thus ψ_{1ij} is the average achievement across both $\text{MSA}_{\text{inf+lit}}$ and RTA for student i in classroom j. The second scale (contrast_{mij}) considers the contrast of the two, that is, RTA performance relative to $\text{MSA}_{\text{inf+lit}}$ performance. Here, contrast_{mij} is an indicator variable coded $1/n_{\text{m}}$ for each of the five subscores of the RTA and $-1/n_{\text{m}}$ for each of the two subscales of the MSA, where n_{m} equals the number of subscores in each scale and thus ψ_{2ij} is the contrast between performance on the RTA versus the $\text{MSA}_{\text{inf+lit}}$ for student i in classroom j. For students with an equal relative ranking on each test this ratio will be one; for students whose relative performance on the RTA is better than their relative ranking on the $\text{MSA}_{\text{inf+lit}}$ this ratio will be positive. For the contrast scale we are seeking to examine whether students' performance on the RTA relative to the $\text{MSA}_{\text{inf+lit}}$ varies systematically between classrooms and whether we can predict this variation.

Appendix C

Annotated Sample Student Assessments

Sample A (rated 1 on all criteria)
"Rats on the Roof" prompt

Yes because is he was lock ebery over and He sead They've got RATS in here! Somebody do something! Call 911! Ooh! Ugh! That haw he isaverd. They wrote a letter like theys To whom I May Concern owng to the excessively noisy disruption of our daytime sleeping schedule. We find ourselves no longer able to remain in residence. We are leaving signed by the Rats on the Roof. that naw he is querd them and cate was to exited.

Sample A received a score of 1 on all criteria in response to the "Rats on the Roof" prompt. The student fails to address the prompt by evaluating characteristics of the Tomcat that would make him helpful in a situation. The student lapses into copying straight from the text rather than offering his/her own argument and explanation. Although the text is used, however, it is not in service of any coherent claim. In terms of organization, style, and mechanics, the writing is highly disjointed, heavily copied, and seriously flawed.

Sample B (rated 3 on all criteria)
"Rats on the Roof" prompt

> I would not like the tomcat to help me with any proublems because the tomcat will just want to live in your home. The tomcat will just want to live in your home. The tomcat will charge you 10 dollars a week. He is not good at his job for example the rat problem he just got scared of the rats and was going crazy breaking things, and screaming. Another reason I don't want the tomcat helping me with a problem because even thow the tomcat got the rats out ther house he broke a lot of stuff in it that will take a long time to replace.

Sample B scored a 3 on all criteria. The student addressed the prompt directly, giving a clear opinion with several supporting reasons. The student also recognized the essential point of the story, that the tomcat did in fact chase the rats away. He/she does not automatically assume the "problem" the cat would be brought in to solve is rats on the roof (which is characteristic of lower-scoring responses); however, he/she does not go beyond the concrete details of the cat given in the story. The student does not infer further about the characteristics of the cat nor offer precise adjectives to describe the cat. The student uses a few good vocabulary words ("charge," "replace") and commits minimal errors in spelling and usage.

> Sample C (rated 2 on all criteria)
> "Millennium Villages" prompt
>
> I think the author think it's for kids in the United States learn about how life is in Kenya I think kids in the United States learn something because life is in Kenya. I think kids in the United States learn something because life in Kenya was very hard because they did not have doctors, no medicine, and now water and electricity. One detail is from the story Kenya is there was 20,000 kids that died because they got bit and there was no medicine. Another detail could be is in Kenya that they had to send the kids to get water and wood so they won't be thirsty. Proverty is poor, having little or no money and that there were people that was like that, they didn't have no food, no water, and no wood to stay warm. There were some people who couldn't afford to get there child into school, so they had to work and get wood and water to survive and other kids could afford to get into a school, they had to wait. Farmers did not have irrigation and fertilizer for there crops. Bed nets are used in every sleeping site in Sauri. The progress is encouraging to supporters to the Millennium Villages project. There are many solutions to the problem that keeps people improverished. Women in Kenya still sit on the ground to sell bananas. Irigation is process by which water is ditrubuted to crops.

Sample C, in response to "Millennium Villages" prompt, received a score of 2 on all criteria. Despite the length of this piece of writing, the content is thin. The response offers the basic observation that "life in Kenya was very hard." In terms of evidence, the student provided more than three pieces demonstrating understanding of life in Sauri before the Millennium Villages project; however, no evidence is provided of how life in Kenya has improved. The organization of the writing reflects a limited sense of beginning, middle, and end. It opens with a series of false-start sentences. The ending is abrupt and, in fact, features sentences copied straight from the text. The middle contains ideas that do not flow. For example, the definition of poverty is

inserted in the middle of the paragraph. As such, there are serious organizational problems. Stylistically, the writing is composed of sentences that are basic in structure and repetitive (e.g., "One detail is . . . . Another detail could be . . . ." Very few connectives are used to join ideas or create flow, and no tier 2 vocabulary is used. Finally, the response is characterized by frequent grammatical errors and awkwardness that impede readers from clearly understanding the writer's meaning (e.g., "One detail is from the story Kenya is there was . . . .").

Sample D (rated 4 on all criteria)
"Millennium Villages" prompt

Kids in the United States should know about Kenya and the Millenium Project so that they can think twice before throwing a temper tantrum when they don't get a toy or candy. A day in Sauri can change their lives.

Sauri, Kenya is home to millions of people who struggle to survive each year. Kids in America don't know how hard it is to battle disease-ridden mosquitoes each night, or trying to find food, or earning enough money to attend school. Every day, 20,000 kids die from Malaria because they can't afford a $5 bednet. Malaria is a disease spread by mosquitoes at night that causes children to die quickly, and adults get very sick.

Malaria is very preventable and treatable. But the people of Sauri can't get the medical attention they need because there are no doctors, medicine, or running water. 3 children share a bed and 2 adults share one. There just isn't enough space for the millions of patients coming in everyday who can't afford health-care.

Students cannot attend the only school in their area, The Bar Sauri Primary School, because their parents can't afford school fees. The students who do go to school have to work hard with the short supply of textbooks all day without lunch. Students can't even eat!

All of this stuff happened 5 years ago, before the Millenium Villages Project. Now, the hospital has enough medicine to treat common diseases and running water. Bed nets are used in every sleeping site in Sauri to prevent Malaria. School now serves a midday meal and they have enough supplies for every student. Keeping people impoverished is possible.

So the next time you want to cry over spilled milk, whine over not having a toy, or throw a temper tantrum over not getting candy, think about the people of Sauri who, for just too long, couldn't even afford a $5 bed net or a good pair of clothes. Don't ever think the world revolves around you, because it doesn't, and millions of kids would give anything to trade places with you.

Sample D, in response to "Millennium Villages" prompt, received a score of 4 on all criteria. The student articulated a plausible reason for why the author may have written the article for kids in the United States and maintains this focus throughout with precise examples from the text. The multiparagraph response is logically organized, with a clear beginning and end that cohere. There are also instances of strong vocabulary and phrases ("disease-ridden," "short supply," "throw a temper tantrum," "revolve"). Although the writing contains some colloquial language ("stuff"), a misused word ("impoverished"), and spelling errors ("Millenium"), such weaknesses are seldom and minor.

Notes

The research reported here was funded through grants from the W. T. Grant Foundation and the Spencer Foundation. The opinions expressed in this article are those of the authors, not the sponsors. The authors remain responsible for any errors in the work.

1. Additionally, specificity has been a distinguishing characteristic for demonstrating effects in meta-analyses of research on professional development (Kennedy, 1999) and in examining instructional interventions (Crandall, Eiseman, & Louis, 1986).

2. Translating lexile measures into "grade-appropriate" designations can, at best, produce rough approximations. In our pilot work, we heard anecdotally from many teachers and literacy coordinators that most of the texts we had chosen were too difficult for most students. The lexile analyzer we used in our post-hoc analysis (obtained from the Lexile Framework for Reading website) suggests that our grade 4 literary text is on par with what students currently read at the middle of the year in grade 4, but below the reading level for suggested CCSS texts. However, the text for grade 5 and 6 students is on par with what students in grades 5–8 currently read and also on par with suggested CCSS texts for grades 5 and 6 (see http://lexile.com/about-lexile/grade-equivalent/grade-equivalent-chart/).

3. The responses were normally distributed, with 2 teachers indicating this occurred monthly, 4 teachers reporting 2 or 3 times a month, 15 reporting it occurred weekly, 6 reporting it happening 2 or 3 times a week, and 2 indicating integration of writing occurring daily.

4. Our scoring criteria parallel standards in the Writing and Language strands. For example, the Academic Language descriptor aligns with the Common Core standard requiring students to "acquire and use accurately grade appropriate general academic and domain-specific words and phrases" (CCSSI, 2010, p. 29).

5. Cohen's kappa (.66 and .68) was also similar for the two prompts demonstrating no differences in agreement between the prompts.

6. A limited set of covariates was imputed due to the sparseness of the covariance matrix with only 18 cases.

7. We use this as a proxy for the expectations teachers hold for student work. While the amount of writing students produce may be theoretically related to their ability to write, we found no empirical evidence for this. Instead, the average number of words is highly related to the percentage of assignments that do not draw directly from worksheets or brief constructed responses, suggesting it is a proxy for teachers' normative expectations.

8. Proportion reduction in variance was calculated using the following formula: $(\tau_{\beta\text{onull}} - \tau_{\beta\text{obackground}})/\tau_{\beta\text{onull}}$.

9. Intraclass correlation was calculated using the following formula: $\tau_{\beta\text{obackground}}/(\tau_{\beta\text{obackground}} + \sigma^2_{\text{background}})$.

10. The percent variance explained was calculated using the following formula: $(\tau_{\beta\text{obackground}} - \tau_{\beta\text{oinstruction}})/\tau_{\beta\text{obackground}}$.

11. However, even after removing the adjustment for prior achievement, the covariate for reasoning and extended writing still did not reach a level of statistical significance ($p = .114$).

12. The samples of student writing in Appendix B are thus the exception rather than the rule since they were specifically chosen to be benchmark assessments because students' scores did not vary on the five criteria.

References

Abrami, P., Bernard, R., Borokhovski, E., Wade, A., Surkes, M., Tamim, R., & Zhang, D. (2008). Instructional interventions affecting critical thinking skills and dispositions: A stage 1 meta-analysis. *Review of Educational Research*, **78**, 1102–1134.

Allensworth, E., Correa, M., & Ponisciak, S. (2008). *From high-school to the future: ACT preparation—too much, too late.* Chicago: Consortium on Chicago School Research.

American Diploma Project. (2004). *Ready or not? Creating a high school diploma that counts*. Washington, DC: Achieve.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Applebee, A. N., & Langer, J. A. (2009). What is happening in the teaching of writing? *English Journal*, **98**(5), 18–28.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., . . . Shepard, L. (2010). *Problems with the use of student test scores to evaluate teachers.* Washington, DC: Economic Policy Institute.

Barr, R., & Dreeben, R. (1983). *How schools work.* Chicago: University of Chicago Press.

Beck, I. L., & McKeown, M. G. (2001). Text talk: Capturing the benefits of read-aloud experiences for young children. *The Reading Teacher*, **55**, 10–20.

Beck, I. L., & McKeown, M. G. (2006). *Improving comprehension with questioning the author: A fresh and expanded view of a powerful approach.* New York: Scholastic.

Bereiter, C., & Scardamalia, M. (1987). An attainable version of high literacy: Approaches to reaching higher-order skills in reading and writing. *Curriculum Inquiry*, **17**(1), 9–30.

Berends, M., Bodilly, S., & Kirby, S. (Eds.). (2002). *Facing the challenges of whole school reform: New American schools after a decade.* Santa Monica, CA: RAND.

Berliner, D. C. (1981). Academic learning time and reading achievement. In J. T. Guthrie (Ed.), *Comprehension and teaching: Research reviews* (pp. 203–226). Newark, DE: International Reading Association.

Brown, R. G. (1991). *Schools of thought: How the politics of literacy shape thinking in the classroom.* San Francisco: Jossey-Bass.

Caro, F. (1971). Issues in the evaluation of social programs. *Review of Educational Research*, **41**(2), 87–114.

Carlisle, J., Kelcey, B., Beribitsky, D., & Phelps, G. (2011). Embracing the complexity of instruction: A study of the effects of teachers' instruction on students' reading comprehension. *Scientific Studies of Reading*, **15**(5), 409–439.

Cohen, D., & Hill, H. (2001). *Learning policy: When state education reform works.* New Haven, CT: Yale University Press.

Common Core State Standards Initiative (CCSSI). (2010). *Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects.* Retrieved from http://www.corestandards.org/assets/CCSSI_ELA%20Standards.pdf

Cooley, W. W., & Leinhardt, G. (1980). The instructional dimensions study. *Educational Evaluation and Policy Analysis*, **2**, 7–25.

Correnti, R., Matsumura, L. C., Hamilton, L. H., & Wang, E. (2012). Combining multiple measures of students' opportunities to develop analytic text-based writing. *Educational Assessment*, **17**(2–3, special issue on measuring instruction), 132–161.

Correnti, R., & Rowan, B. (2007). Opening up the black box: Literacy instruction in schools participating in three comprehensive school reform programs. *American Educational Research Journal*, **44**, 298–338.

Crandall, D., Eiseman, J., & Louis, K. (1986). Strategic planning issues that bear on the success of school improvement efforts. *Educational Administration Quarterly*, **22**(3), 21–53.

Crosson, A. C., Matsumura, L. C., Correnti, R., & Arlotta-Guerrero, A. (2012). The quality of writing tasks and students' use of academic language in Spanish. *Elementary School Journal*, **112**, 469–496.

D'Agostino, J. V., Welsh, M. S., & Corson, N. M. (2007). Instructional sensitivity of a state's standards-based assessment. *Educational Assessment*, **12**, 1–22.

Diamond, K. E., & Powell, D. R. (2011). An iterative process to the development of a professional development intervention for Head Start teachers. *Journal of Early Intervention*, **33**(1), 75–93. doi:10.1177/1053815111400416

Hauck, K., & Street, A. (2006). Performance assessment in the context of multiple objectives: A multivariate multilevel analysis. *Journal of Health Economics*, **25**, 1029–1048.

Hiebert, J., Gallimore, R., & Stigler, J. W. (2002). A knowledge base for the teaching profession: What would it look like and how can we get one? *Educational Researcher*, **31**(5), 3–15.

Hirsch, E. D., & Pondiscio, R. (2010). There's no such thing as a reading test. *The American Prospect.* Retrieved from http://prospect.org/article/theres-no-such-thing-reading-test

Hoffman, L., & Rovine, M. (2007). Multilevel models for the experimental psychologist: Foundations and illustrative examples. *Behavior Research Methods*, **39**, 101–117.

Jeffery, J. V. (2009). Constructs of writing proficiency in U.S. state and national writing assessments: Exploring variability. *Assessing Writing*, **14**, 3–24.

Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.

Kennedy, M. (1999). Form and substance in mathematics and science professional development. *National Institute for Science Education Brief*, **3**(2), 1–7.

Knapp, M., & Associates. (1995). *Teaching for meaning in high poverty classrooms.* New York: Teachers College Press.

Leinhardt, G., & Seewald, A. M. (1981). Overlap: What's tested, what's taught? *Journal of Educational Measurement*, **18**(2), 85–96.

Maryland State Department of Education. (2012). *2010 MSA Reading technical report.* Retrieved from http://www.marylandpublicschools.org/MSDE/divisions/planningresultstest/2010+MSA+Reading+Technical+Report.htm

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, **23**(2), 13–23.

Murphy, P. K., Wilkinson, I. A. G., Soter, A. O., Hennessey, M. N., & Alexander, J. F. (2009). Examining the effects of classroom discussion on students' comprehension of text: A meta-analysis. *Journal of Educational Psychology*, **101**(3), 740–764.

Muthén, L., & Muthén, B. (2010). *MPlus statistical analysis with latent variables: User's guide*. Los Angeles: Muthén & Muthén.

National Commission on Writing. (2003). *The neglected "R": The need for a writing revolution*. New York: College Board.

Nickerson, R. (1989). On improving thinking through instruction. *Review of Research in Education*, **15**, 3–57.

Nystrand, M. (2006). Research on the role of classroom discourse as it affects reading comprehension. *Research in the Teaching of English*, **40**, 392–412.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, **4**, 525–556.

Raudenbush, S. (2007). Designing field trials of educational innovations. In B. Schneider & S. K. McDonald (Eds.), *Scale-up in education: Issues in practice*. New York: Rowman and Littlefield.

Raudenbush, S. W., & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.

Raudenbush, S., Rowan, B., & Cheong, Y. (1993). Higher order instructional goals in secondary schools: Class, teacher and school influences. *American Educational Research Journal*, **30**, 523–553.

Raudenbush, S., Rowan, B., & Kang, S. J. (1991). A multilevel, multivariate model for studying school climate with estimation via the EM algorithm and application to U.S. high-school data. *Journal of Educational and Behavioral Statistics*, **16**, 295–330.

Resnick, L. B. (1987). *Education and learning to think.* Washington, DC: National Academy Press.

Resnick, L. B. (2010). Nested learning systems for the thinking curriculum. *Educational Researcher*, **39**, 183–197.

Resnick, L. B., & Klopfer, L. E. (Eds.). (1989). *Toward the thinking curriculum: Current cognitive research.* Alexandria, VA: Association for Supervision and Curriculum Development.

Rosenshine, B., & Furst, N. (1973). The use of direct observation to study teaching. In R. M. W. Travers (Ed.), *Second handbook of research on teaching* (pp. 122–183). Chicago: Rand McNally.

Rowan, B., Camburn, E., & Correnti, R. (2004). Using teacher logs to measure the enacted curriculum: A study of literacy teaching in third-grade classrooms. *Elementary School Journal*, **105**, 75–102.

Ruiz-Primo, A., Shavelson, R., Hamilton, L. S., & Klein, S. P. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal for Research in Science Teaching*, **39**, 369–393.

Schmidt, W., McKnight, C., Houang, R., Wang, H., Wiley, D., Cogan, L., & Wolfe, L. (2001). *Why schools matter: A cross-national comparison of curriculum and learning*. San Francisco: Jossey-Bass.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* Thousand Oaks, CA: Sage.

Stebbins, L., St. Pierre, R., Proper, E., Anderson, R., & Cerva, T. (1977). *Education as experimentation: A planned variation model: Vol. IV-A. An evaluation of follow-through.* Cambridge, MA: ABT.

Thum, Y. M. (1997). Hierarchical linear models for multivariate outcomes. *Journal of Educational and Behavioral Statistics*, **22**, 77–108.