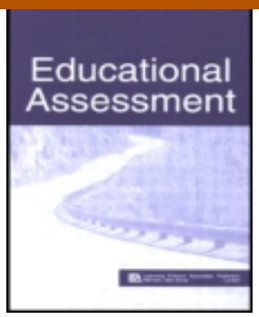

Educational Assessment

ISSN: 1062-7197 (Print) 1532-6977 (Online) Journal homepage: http://www.tandfonline.com/loi/heda20

Combining Multiple Measures of Students' Opportunities to Develop Analytic, Text-Based Writing Skills

Richard Correnti , Lindsay Clare Matsumura , Laura S. Hamilton & Elaine Wang

To cite this article: Richard Correnti , Lindsay Clare Matsumura , Laura S. Hamilton & Elaine Wang (2012) Combining Multiple Measures of Students' Opportunities to Develop Analytic, Text-Based Writing Skills, Educational Assessment, 17:2-3, 132-161, DOI: 10.1080/10627197.2012.717035

To link to this article: http://dx.doi.org/10.1080/10627197.2012.717035

Educational Assessment, 17:132–161, 2012

ISSN: 1062-7197 print/1532-6977 online
DOI: 10.1080/10627197.2012.717035

Routledge
Taylor & Francis Group

Combining Multiple Measures of Students' Opportunities to Develop Analytic, Text-Based Writing Skills

Richard Correnti and Lindsay Clare Matsumura
University of Pittsburgh

Laura S. Hamilton
RAND Corporation

Elaine Wang
University of Pittsburgh

Guided by evidence that teachers contribute to student achievement outcomes, researchers have been reexamining how to study instruction and the classroom opportunities teachers create for students. We describe our experience measuring students' opportunities to develop analytic, text-based writing skills. Utilizing multiple methods of data collection—writing assignment tasks, daily logs, and an annual survey—we generated a composite that was used in prediction models to examine multivariate outcomes, including scores on a state accountability test and a project-developed response-to-text assessment. Our findings demonstrate that students' opportunities to develop analytic, text-based writing skills predicted classroom performance on the project-developed response-to-text assessment. We discuss the importance of considering the measure(s) of learning when examining teaching–learning associations as well as implications for combining multiple measures for purposes of better construct representation.

Instructional quality has risen to the top of the country's education reform agenda as a result of research showing the critical role teaching plays in student achievement (Rowan, Correnti, & Miller, 2002; Sanders & Rivers, 1996). Indeed, variation among teachers has been demonstrated to be the most important factor within the control of schools that influences students' learning outcomes (Nye, Konstantopoulos, & Hedges, 2004; Rivkin, Hanushek, & Kain, 2005; Sanders, Wright, & Horn, 1997), and this variation is likely to occur in large part as a result of differences in the instructional activities to which students are exposed. However, efforts by researchers

Correspondence should be sent to Richard Correnti, 803 Learning Research and Development Center, University of Pittsburgh, 3939 O'Hara Street, Pittsburgh, PA 15260. E-mail: rcorrent@pitt.edu

and policymakers to identify the teacher and classroom characteristics that are associated with student learning have been only minimally successful, and there is broad recognition of a need for new measures of instructional quality that can be administered in large numbers of classrooms for the purpose of documenting the practices of effective teachers.

The quest to understand what distinguishes high- and low-quality instruction is not new, of course. For decades, researchers have sought to document generalizable aspects of teaching that influence student learning and that can be reproduced (see, e.g., Brophy & Good, 1986; Dunkin & Biddle, 1974; Medley, 1977). Some efforts to assess instructional quality occurred in the context of a broader research agenda that sought to evaluate and promote equity in students' learning opportunities beginning in the 1960s. The idea of measuring "opportunity to learn" gained new traction in the late 1980s and early 1990s when standards-based accountability policies were being debated at the federal and state levels. For instance, the National Council on Education Standards and Testing (NCEST), a group that was convened to advise government officials on whether and how to create a national system of standards and assessments, argued for the inclusion of *school delivery standards* in a broader standards-based accountability system (see, e.g., Porter, 1993, 1995). These standards, later called "opportunity-to-learn standards," were intended to ensure that students had access to the resources (including materials, instructional practices, and school conditions) they would need in order to demonstrate mastery of content standards (Jennings, 1998; NCEST, 1992).

The conceptualization of Opportunity-to-Learn (OTL) has evolved over time. Early generations of research on students' OTL (e.g., cross-national studies conducted by the International Association for the Evaluation of Educational Achievement) relied heavily on large-scale surveys that focused on time-on-task or content that overlapped with the tasks and content represented on student assessments (see, e.g., Husen, 1974; McKnight et al., 1987; Schmidt & McKnight, 1995). Over time, the definition of OTL expanded beyond theories of time on content (see, e.g., Carroll, 1963) to include multiple areas of practice associated with student learning outcomes. This more comprehensive vision of OTL included a focus on not just the content that was covered in classrooms but also what teachers do in the classroom, the activities in which students engage, and the materials and other resources that are used to support instruction (Brewer & Stasz, 1996; Herman, Klein, & Abedi, 2000). These areas of practice are not mutually exclusive, but they provide a helpful framework for considering the range of components of classroom instruction that might be included under the OTL umbrella.

The multidimensional nature of OTL, however, presents many measurement challenges (Baker, 2007). Further complicating the problem of measurement is subject-matter differences in what constitutes high-quality instruction. Research suggests that teaching is not a generic practice but is mediated by the subject-matter content (Grossman, Stodolsky, & Knapp, 2004; Stodolsky & Grossman, 1995). What constitutes effective teaching, therefore, may vary across different areas of the curricula. Researchers and others who seek to understand OTL must make decisions about what particular subject-matter content to target and then identify or develop appropriate measures for that content. For example, although Carlisle, Kelcey, Beribitsky, and Phelps (2011) formulated and simultaneously examined three different constructs of teacher–student behavior, they examined these constructs within the context of lessons focused on reading comprehension instruction rather than capture the full range of the literacy curriculum. For studies that pose a specific question about a single aspect of teaching, such as teachers' knowledge of teaching a particular subject (Hill, Rowan, & Ball, 2005), the use of a single

measurement approach (e.g., surveys) might be appropriate. However, if researchers are instead trying to understand the extent to which students are exposed to a broader array of opportunities to learn particular subject-matter content, multiple dimensions of practice will need to be examined within the broader theme of OTL and, perhaps, multiple measures will also be necessary.

The goal of our study is to investigate an approach to combining multiple measures to assess students' OTL within a particular domain of the language arts curriculum: analytic text-based writing. Specifically, we draw on both performance- and survey-based approaches (writing tasks, instructional logs, and an annual survey) to create a rich measure of students' analytic text-based writing opportunities in classrooms.

ANALYTIC TEXT-BASED WRITING

We focus on measuring students' opportunity to write analytically in response to text for two key reasons: First, this skill is strongly emphasized in the 2010 Common Core State Standards (CCSS), representing a significant shift from current individual state standards and common practices in schools (Rothman, 2011). Specifically, the standards prioritize students' abilities to read complex texts and adopt an analytic stance in their writing about text by focusing on the "perceived merit and reasonableness of the claims and proofs" presented in a text rather than the "emotions that the writing evokes in the audience" (CCSS Appendix A, 2010, p. 24). We refer to this set of skills as "analytic text-based writing." Although assessments for the CCSS are still in development, it is likely that "reading will be assessed through writing, making writing even more critical" in the curricula (Calkins, Ehrenworth, & Lehman, 2012, p. 10). Measures of classroom practice are needed that assess students' opportunity to achieve the learning goals set forth in the CCSS, including the development of analytic text-based writing skills.

Second, given the poor state of writing and reading comprehension instruction in schools (Applebee & Langer, 2009; Lee, Grigg, & Donahue, 2007) it is likely that increasing numbers of instructional interventions and reforms will be initiated that are intended to support teachers' skills at teaching to the Common Core State Standards. Ways to measure students' opportunities to develop analytic, text-based writing skills will be needed to help monitor the progress and effectiveness of targeted interventions and reform programs.

Writing Assignment Tasks

Our conceptual framework for measuring students' opportunities to develop their analytic, text-based writing skills focuses on the quality of the classroom tasks students are assigned and the instruction students receive to write and to reason analytically about texts. Numerous studies show a relationship between high-quality classroom writing tasks and increased student learning outcomes (American Institute for Research, 2005; Matsumura, Garnier, Pascal, & Valdés, 2002; Newmann, Bryk, & Nagaoka, 2001). Specifically, research indicates that access to cognitively demanding classroom tasks that guide students to construct knowledge (i.e., interpret, analyze, synthesize, or evaluate information) as opposed to recalling facts and generating surface-level summaries is predictive of students' scores on standardized tests of reading achievement (Matsumura, Garnier, Pascal, & Valdés, 2002; Matsumura, Garnier, Slater, & Boston, 2008;

Newmann et al., 2001). Students' opportunities to engage with classroom writing tasks that guide them to generate elaborated written responses (i.e., support conclusions, generalizations or arguments through extended writing) also is associated with increased standardized test scores (Matsumura, Garnier, et al., 2002; Matsumura, Garnier, et al., 2008; Newmann et al., 2001).

Cognitively demanding classroom writing tasks also have been associated with increased quality of students' writing (American Institute of Research, 2005; Crosson, Matsumura, Correnti, & Arlotta-Guerrero, 2012; Matsumura, Patthey-Chavez, Valdés, & Garnier, 2002). Matsumura, Patthey-Chavez, et al. (2002), for example, found that the cognitive demand of writing tasks predicted a small but significant amount of variation in the content of students' written work. Crosson et al. (2012) similarly showed that the cognitive demand of writing tasks predicted the overall quality of the content of students' writing in their native language (Spanish) as well as students' use of most features of academic language, including academic vocabulary, embedded clauses, temporal and causal connectives, and use of a variety of connectives.

Instructional Strategies

Although a fair amount of research has linked the quality of classroom tasks to students' academic outcomes, integrated theories of how instructional strategies combine with high cognitive demand tasks are currently lacking (Benko, 2012). Guided by the extant research on effective literacy instruction, we propose that two broad areas of instructional practice likely contribute to developing students' analytic, text-based writing skills. The first area of instructional practice is informed by theory and research that foregrounds the importance of providing students with general knowledge of writing, such as an understanding of the writing process and writing strategies (Calkins, 1986; Graham & Perin, 2007). Graham and Harris (1993), for example, found that developing students' self-awareness of the skills involved in composing by providing them with step-by-step strategies for planning, drafting, and revising their work can lead to significant increases in the quality of students' writing. In this tradition, arguments for more time on writing in the classroom include that it manifests higher cognitive outcomes in reading comprehension (Morrow, 1992), that writing is valued because it is more metacognitively challenging (Langer & Applebee, 1987; Tierney, Soter, O'Flahavan, & McGinley, 1989), and that providing more writing instruction is characteristic of excellent literacy teachers (Knapp et al., 1995; Pressley, Allington, Wharton-McDonald, Block, & Morrow, 2001).

The second area of practice is informed by theory and research that suggests that focusing purely on the form and process of writing is likely insufficient for developing high-level academic writing skills (Graham & Perin, 2007; Hillocks, 2006). Students need instruction focused on mastering particular types of writing in the service of learning (Boscolo & Carotti, 2003; Hillocks, 2006; Shanahan & Shanahan, 2008). From this perspective, classroom instruction—activities and discussions—should integrate the application of higher level analytic comprehension skills with writing that reflects these critical thinking processes (Hillocks, 2006). This is especially the case after the primary grades when writing begins to play an increasing role in students' mastery of academic content (Hillocks, 1984, 2006).

Indeed, research suggests that different kinds of writing activities can lead to different learning outcomes (Graham & Perrin, 2007; Klein, 1999; Langer & Applebee, 1987; Marshall, 1987). Analytic writing, specifically, may promote more "complex and thoughtful inquiry" of targeted academic content than constrained, short-answer responses that break information

into small pieces (Langer & Applebee, 1987, pp. 135–136). Boscolo and Carroti (2003), for example, compared two approaches to teaching literature to high school students: analytic writing as a tool to understanding the course content and a traditional approach to literature sponsored in the course. Results indicated that students in the analytic writing condition produced higher quality commentaries on a literary text and were more likely to perceive the usefulness of academic and personal writing.

MEASURING STUDENTS' ANALYTIC TEXT-BASED WRITING SKILLS

Following from our review of the literature, our hypothesis is that the development of students' analytic, text-based writing skills is linked to an opportunity structure consisting of four domains: writing tasks that are of a high cognitive demand, opportunities for students to provide elaborated communication, activities analyzing and synthesizing text in the context of discussing or doing writing (i.e., activities integrating comprehension and writing), and time (repeated exposure) to develop general writing skills (e.g., application of the writing process). The goal of this study was to develop a measure of this opportunity structure. We used multiple measures generated through multiple methods in order to quantify students' opportunities in these four domains. As shown in Figure 1, our composite measure draws on classroom assignment tasks, instructional logs, and an annual survey. Each domain is captured, at least partly (dashed arrows represent moderate coverage of the sampled domain), by more than one measure, with teacher logs overlapping with elements of measuring both the task, and how frequently it is taught.

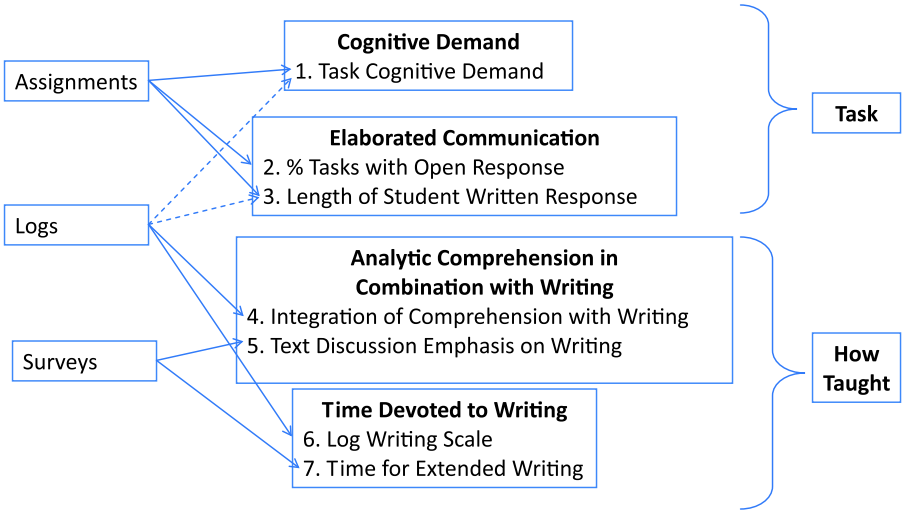

FIGURE 1 Utilizing multiple methods for capturing students' opportunities to develop the skills to form an argument in response to text and communicate their argument in writing. (color figure available online)

Purpose of the Study

In the current study we investigated the value of combining multiple measures to assess the opportunities afforded to students to develop their analytic, text-based writing skills. To provide insight on the technical quality of our composite measure, we began by examining the correlations among our measures and used factor analysis to explore the feasibility of creating a single measure representing teaching behavior(s) that support analytic, text-based writing. Second, we examined the relationship between our measures and student outcomes in order to evaluate the extent to which our measures captured the curricular content and instructional strategies that are associated with student learning. We examined whether findings from a composite measure of students' opportunities-to-learn (OTL) analytic, text-based writing skills would change inferences about students' OTL that we would have made if we considered individual measures by themselves. Specifically, we examined the association between individual component OTL measures and student learning as measured through a state accountability test as well as on a project-developed response-to-text assessment. Because these analyses are based on a small sample of 18 teachers, they should be considered exploratory. The findings do not provide definitive information about how the measures of teaching practice are likely to work on a large scale, but they do suggest ways in which multiple measures could be useful in future research studies and in practice.

Sample

Our data include 426 students nested in 18 classrooms within a single urban district in Maryland. Student-level data included scores on two different assessments administered in spring 2011. Data about teachers' instructional practices were collected throughout the year using the measures described next. Teachers' background characteristics, used for controls in our linear prediction models, were collected on the annual survey. Descriptive statistics for the student and teacher samples are presented in Table 1. The demographic characteristics of students in our sample were roughly representative of the larger district, with the exception that minorities were slightly overrepresented.

Operationalization of Opportunity to Learn Measures

Corresponding to Figure 1, we collected information on classroom teaching with respect to students' opportunities to develop text-based writing. Next we describe each of our individual component measures, grouped by method of data collection.

Classroom assignment tasks with student work. The Instructional Quality Assessment was used to measure the quality of teachers' assignments (Junker et al., 2006; Matsumura, Slater, & Crosson, 2008). Teachers were asked to submit six responses to text-writing tasks that they considered to be challenging. Three of the tasks were collected in November and three were collected in May. For each task, teachers were asked to complete a cover sheet on which they described the task and the text to which the students responded; the directions they gave to students to complete the task; the grading criteria for determining high-, medium-, and low-quality work; and the way that they communicated their expectations to students (e.g.,

TABLE 1
Student and Teacher/Classroom Demographics

	M	*SD*	*Range*
Student level[a]			
Prior reading achievement	0.00	1.00	−2.09 to 3.4
Prior mathematics achievement	0.00	1.00	−3.08 to 2.83
Absences	7.58	6.63	0–40
Free lunch	.45	.50	0–1
Reduced-price lunch	.12	.32	0–1
Black	.80	.40	0–1
Hispanic	.13	.34	0–1
Native American	.11	.32	0–1
Asian	.05	.23	0–1
White	.03	.17	0–1
IEP	.11	.31	0–1
Teacher level[b]			
Years' experience	16.84	10.99	2–38
PhD	.17	.38	0–1
Advanced certification	.17	.38	0–1
Grade	4.72	.67	4th (8), 5th (9), and 6th (1)

[a]$n = 426$. [b]$n = 18$.

in a rubric or class discussion). For each task teachers also submitted four pieces of student work—two they considered to be of high quality and two they considered to be of medium quality for the class. The writing tasks and associated student work were assessed along three dimensions: the cognitive demand of the task, opportunity for open response, and the average length of students' written responses.

The *Task Cognitive Demand* was assessed on a 4-point scale, from 1 (*poor*) to 4 (*excellent*), and rated the degree to which each writing task supported students in applying higher level, analytic thinking skills (as opposed to recalling or identifying basic information from a text) and in using appropriate evidence and details from a text to support their assertions. To receive a score of 4 on this dimension, the writing task would guide students to construct meaning beyond the surface-level events in a text and write extended responses that include details and evidence from the text to support their assertions. To receive a score of 2, the writing task would guide students to construct surface-level summaries of a text (i.e., recall the beginning, middle, and end of a story). Writing tasks that receive the lowest score of 1 guide students to recall isolated, disconnected facts about a text and to provide little or no evidence to support their assertions. Alternatively these tasks may require students to write on a topic that is not connected to the content of a text. Despite the fact that we asked for challenging writing tasks, the mean rating across all teachers ($u = 2.08$, $SD = .39$) demonstrated that the average task required only recall of text. Mean assignment ratings across the 18 teachers ranged from 1.33 to 2.83.

In addition to examining the cognitive demand of the writing tasks, we also considered the opportunity these tasks provided to students for elaborated written communication (Newmann, Lopez, & Bryk, 1998). *Opportunity for Open Response* was calculated as the percentage of writing tasks that did not constrain students' response to only a few sentences (i.e., followed an

open as opposed to a constrained format). On average, 32% of the writing tasks we collected invited open (and extended) responses from students. The proportion of writing tasks each teacher submitted with a constrained format ranged from all to none with a slight negative skew to the distribution. It bears noting that most of the constrained format writing tasks we received resulted directly from a district-led initiative to have students create "brief constructed responses" that mimic what students are asked to do on the state accountability test.

Finally, using the student work we also considered the *Length of Students' Written Responses*. We considered this a proxy for teachers' criteria for task completion (Doyle, 1983), or, in other words, how much students were expected to write to satisfy the requirements of the writing task. To obtain this score, we averaged the number of words produced by students across the four pieces of student work for each of the writing tasks. On average, students wrote about 116 words per assignment, whereas teachers' classroom averages ranged from about 38 words per task to 230 words per task.

Instructional logs. Over the course of data collection in the 1st year, teachers were asked to complete online daily surveys of their literacy teaching practice (instructional logs) at three different points in the academic year—2 weeks in November, 2 weeks in January, and 2 weeks in May (30 days total). Participating teachers turned in nearly complete data ($M = 26$ logs per teacher, $SD = 4.28$), and reported that the logs took about 5 to 10 min to complete ($M = 8.7$). The logs sample literacy teaching across three large domains of elementary language arts instruction, including reading comprehension, writing, and word analysis. Items on the log included those with a focus on writing instruction and on the integration of reading and writing, which are germane for our four sampled domains contributing to students' opportunities for extended written responses. The logs also contained additional items in both comprehension and word analysis. To prepare items for analysis, a data reduction technique was utilized collapsing individual log items into dichotomous indicators of whether a particular aspect of instruction was a focus on that day (for item groupings, see, e.g., Correnti, 2005; Correnti & Rowan, 2007). The frequency of days when these 29 dichotomous practices were covered is described in Table 2.[1]

Using the 29 dichotomous items, we sought to create two scales from the log data corresponding directly and indirectly to our sampled domains. The *Log Writing Scale* measured students' exposure to writing across the sampled days and included all nine items contained in the second column of Table 2. As represented in Figure 1, this was a direct measure of the frequency with which students received exposure to the direct teaching of writing or participated in the writing process across all sampled occasions; it is, thus, representative of classroom time devoted to writing. Indirectly it is also a measure of students' opportunities for elaborated communication because the amount students wrote, and other elements of the task (including more cognitively demanding tasks such as a written literature extension project), are embedded within this measure (Figure 1).

[1]Only slight differences in frequency are noted between this sample of teachers and the much larger sample from the Study of Instructional Improvement using nearly identical language arts logs. The one main difference between the two studies was the type of log administration which was paper and pencil in the Study of Instructional Improvement but online in the current study.

TABLE 2
Frequency (% of Days Strategy Was Covered) of Log Items Used to Construct Scales in Measurement Models

Comprehension	%	*Writing*	%	*Word Analysis*	%
Check understanding	51	Teacher directed	31	Assess student reading	14
Student discussion	48	Prewriting	24	Focus on comprehension same day	14
Brief answers	47	Teacher comments	20	Teacher directed	12
Teacher directed	47	Writing practice	17	Context picture cues	09
Prereading	44	Write paragraphs[a]	17	Focus on write same day	09
Story structure	43	Integrate comprehension[a]	16	Structural analysis	06
Analyze/Evaluate[a]	33	Literary techniques[a]	11	Sight words	04
Focus on writing same day[a]	26	Substantive revisions[a]	10	Phonics cues	03
Integrate writing[a]	23	Edit	08	Letter-sound relations	03
Extended answers[a]	20				

[a]Item is part of literature-based scale that is contrasted in measurement model with all other log items.

The *Integration of Comprehension with Writing Scale* is a measure of the ratio of time teachers spent integrating comprehension and writing in literature-based activities versus other literacy content (see Table 2; bold items represent literature-based items and they are contrasted against all other items). This measure has two functions that may be theoretically meaningful. First, it is a relative measure of the proportion of time spent in literature-based activities rather than simply totaling up the frequency count of items covered. The relationship of literature-based topics to other content may be important in distinguishing between teachers, especially where teachers differ in their teaching patterns (e.g., someone whose content coverage is characterized by depth vs. breadth). Second, these items represent less frequently taught topics. Thus, in addition to being a direct measure of the integration of writing with reading comprehension (including such items as examining literary techniques within comprehension, writing extended answers to comprehension questions, or working on a written literature extension project), it is also an indirect measure of cognitive demand because it represents the ratio of higher order content to other content (see Figure 1). Such a ratio can be a variable of interest for considerations of equity (Gamoran, Porter, Smithson, & White, 1997) and can help identify associations between teaching and learning (Correnti, Phelps, & Kisa, 2010).

Computation of these scales was conducted using the following procedures. To account for the multidimensional aspects of instruction, we created a multilevel multivariate measurement model using HLM7.0 (Raudenbush & Bryk, 2002). Items for a given day were nested in days, and days were nested in teachers. These models are useful for understanding psychometric properties of the scales (see, e.g., Raudenbush, Rowan, & Kang, 1991), adjusting for covariates in the model (e.g., time of the year and item characteristics), and computing an empirical Bayes residual to be used as a continuous indicator of the frequency of students' opportunities to learn (see, e.g., Carlisle et al., 2011). The measurement model indicated significant variation existed between teachers on both the Log Writing Scale, $\chi^2(12) = 75.54$, $p < .000$, and the Integration of Comprehension With Writing Scale, $\chi^2(12) = 53.35$, $p < .000$. Teacher-level reliabilities were .67 and .78, respectively, for the two scales. The Log Writing Scale, consisting of the teacher-level empirical Bayes residuals resulting from the measurement model, was roughly

normally distributed ($\mu = 0$, $SD = 1$) with a slight negative skew because the modal score was below 0. Teachers in the bottom third of the distribution[2] taught writing less frequently (25% of all occasions vs. 38% and 45% for middle and upper[3] tiers, respectively) and also were less likely to focus on comprehension and writing on the same day (16% of all occasions vs. 24% and 39%, respectively). These descriptive statistics indicate between-classroom variation in the frequency writing was incorporated in teachers' language arts instruction. In general, writing was taught in slightly more than one third of all lessons (36%), whereas reading comprehension was an emphasis in about two thirds of the lessons (66%).

Surveys. Teachers participated in a survey at the end of the academic year. One item stem examined teachers' self reports of practices they incorporated in their text discussions. We created a four-item measure focused on the frequency teachers reported engaging in activities related to Text Discussion with an Emphasis on Writing. Items were answered on a scale of 1 (*never*) to 5 (*almost always*). The items included (a) students identify the author's purpose, (b) students discuss elements of the writer's craft, (c) students make connections between ideas/literary elements within or across texts, and (d) students analyze and evaluate each other's assertions. A one-factor solution was obtained, which explained 63% of the variance in the items. The items were revealed to have strong internal consistency with a Cronbach's alpha of .80. Higher scores on this factor indicate a tendency in text discussions to discuss literary elements and explore the writer's craft in the context of making connections and critically analyzing the text. The average score of 3.2 indicates that students engage in these activities in their class discussions closer to "sometimes" (a score of 3 on the scale) than they do "often" (a score of 4 on the scale). This mean ranked lowest among teachers' self-reported instructional behaviors on the survey, indicating that this type of instruction occurred less frequently than other surveyed items.

A second measure was created from the annual survey where teachers were asked to describe how they distributed their instructional time across a list of multiple activities. For example, they were asked to estimate the proportion of time students spent in Language Arts instruction that was devoted to (a) reading text and improving reading skills (e.g., independent reading), (b) text discussion activities (e.g., identifying main ideas, infer meaning for text), (c) writing (e.g., practice writing mechanics, learn writing elements, write in response to open-ended prompts or to text they had read), and (d) assessing students' understanding (e.g., multiple-choice/fill-in-the-blank questions, practicing for the state accountability test). Elsewhere they were also asked about the proportion of time students spent writing for the following purposes: (a) to gain practice, (b) to create written expositions from own ideas, (c) to provide written responses to text primarily to summarize, and (d) to produce extended written responses analyzing and evaluating texts. To generate a measure of Time for Extended Writing we created a product utilizing these two questions. Time for extended writing was produced by multiplying the proportion of time spent writing in the first question and the proportion of writing time they

[2]We created cut points around our standardized writing scale such that the bottom tier consists of teachers whose scores were between $^1/_2$ standard deviation below the mean and $1^1/_2$ standard deviations below the mean, and the top tier consists of teachers whose scores were between $^1/_2$ standard deviation above the mean and $1^1/_2$ standard deviations above the mean; and the middle tier were in between the two.

[3]Omits highest scoring teacher because their score was higher than $1^1/_2$ standard deviations above the mean.

spent having students produce extended written responses analyzing and evaluating texts. On average teachers' self-reports indicate that they spend proportionally little of their time (a little more than 4%) having students write extended responses analyzing or evaluating texts.

Handling of missing data. Teachers participated in data collection efforts that spanned the course of the year. Because data collection spanned different methods, including daily literacy logs, response-to-literature assignments, and a spring survey, multiple opportunities existed for teacher nonresponse. Sixteen of the 18 teachers had data from all three sources, whereas two teachers had task and log data but were missing the annual survey. Rather than use listwise deletion to remove cases without complete data on instruction, we adopted a strategy of multiple imputation for participants with incomplete data. We conducted a two-level multiple imputation using MPlus6.12 (Muthen & Muthen, 1998–2010) to generate five data sets with complete data. Our analytic procedures utilized all five data sets in our analyses (methods used for calculating estimates and standard errors are detailed in Peugh & Enders, 2004).

Student Learning Outcomes

A distinct feature of our study was the commitment to understanding student learning via multiple measures. We examined student learning on both the state accountability test—the Maryland School Assessment (MSA)—and on a project developed Response-to-Text Assessment (RTA). Next we describe each assessment and briefly discuss how preliminary analyses of the relationship between outcome measures suggest the utility of multivariate models to test the predictive validity of measures of students' opportunities to develop analytical text-based writing.

Maryland School Assessment. The MSA is a large-scale standardized assessment that measures students' progress toward attaining the reading skills specified in the state's curriculum. The reading 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" (MSDE, 2012). Students must respond to each prompt within the eight lines provided. In terms of scoring, the BCR is given a rating of 0 to 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 items and two BCRs on each scale). The test publisher created scale scores for each subscale and for the test overall.

Response to Text Assessment. The RTA is designed to assess students' ability to write analytically in response to text, use appropriate evidence from a text to support their claims, and apply other features of academic writing (e.g., language use, mechanics). The RTA was administered by teachers in May at the end of the academic year. In the fourth, fifth, and sixth grades, the classroom teacher read a text aloud to students who followed along with their own copies. Teachers stopped at predetermined places to check students' understanding and clarify

vocabulary that researchers posited might be unfamiliar to students (e.g., hasty, irrigation). Students also were encouraged to underline the text and take notes as they read.

In the fourth grade, students responded to a short story by James Marshall (*Rats on the Roof*) about a pair of dogs that enlist a cat to help them solve their rat problems. The cat solves the problem but, ironically, not in the way the dogs intended. The prompt students responded to was, "Is the Tomcat someone you would want to help you solve a problem? Why or why not? Use at least three or four examples from the text to explain your answer." In the fifth and sixth grades, students responded to a feature article from *Time for Kids* about a United Nations–supported effort to eradicate poverty in a rural village in Kenya. Students then responded to the following prompt: "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 three examples of what life in Kenya was like before the *Millennium Villages* project and what life is like now."

Students' responses were scored on five dimensions each of which was assessed on a 4-point scale, from 1 (*low*) to 4 (*excellent*). *Evaluation* 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 students for varied sentence lengths and complex syntactical structures, multiple uses of tier-two 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* reflected their ability to adhere to grade-appropriate standard writing conventions.

Students' responses were coded by a member of the research team. To check the reliability of the scores, a second member of the team coded 20% of the responses selected at random at each grade level 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 scores assigned by the two raters. It showed the exact match between raters was 79% with only two instances of raters differing by more than one. Cohen's kappa (.672), $\chi^2(9) = 603.94$, and the Pearson correlation ($r = .828$) both indicate moderately high agreement between raters overall.

Correlations and factor structure of the MSA and RTA. We sought to understand the extent to which the RTA measured the specific skill of responding analytically in writing to a prompt based on a text students had just had the chance to comprehend. Correlations among each of the five dimensions of the RTA rubric with the MSA scale score range from .34 to .51 at the student level and connote a statistically significant association between the MSA and RTA. However, the moderate correlations also suggest that the abilities assessed by the two assessments do not overlap completely. 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.

Using SPSS19.0, we conducted factor analyses of the subscales of each achievement outcome (the MSA had three subscales, whereas the RTA had five scoring dimensions) using Principal Axis Factoring with an Oblique rotation. A single factor with eigenvalue greater

than 1 was extracted for the RTA and MSA when subscores for each assessment were entered separately. We were also interested in the dimensionality of the subscores of the two assessments together. In this analysis, two factors were extracted with eigenvalues greater than 1. The scree plot of the eigenvalues confirms a linear slope after the second factor. We examined both the pattern and structure matrix, which demonstrated that the five subscores of the RTA loaded more highly on the first factor, whereas the three subscores of the MSA loaded more highly on the second factor.

Although parsimony favors the simplest solution—a single unidimensional latent ability construct—there is also evidence of a second factor contrasting performance on the RTA with student performance on the MSA. Therefore, we decided to explore two different multilevel multivariate models. The first of these examined the RTA and MSA as separate outcomes simultaneously. This allowed us to examine instructional covariates on each scale while accounting for the covariance between the measures. The second multivariate model examined a unidimensional latent achievement score as one outcome while also modeling the contrast between students' performance on the RTA relative to performance on the MSA as a second outcome. When these multivariate models are considered simultaneously, they help the reader interpret statistically significant findings of covariates on the contrast (Raudenbush et al., 1991).

ANALYSES

We examined the correlation matrix of our operationalized measures for purposes of triangulation where we expected to see convergence among our measures. We then conducted a factor analysis to examine the underlying structure of our seven individual components composing the opportunity structure provided to students. We then formed a composite measure representing students' opportunities to develop proficiency at producing analytical, text-based writing. Subsequently we examined the predictive validity of analytical, text-based writing in a series of multilevel multivariate models using HLM7.0 (Raundenbush & Bryk, 2002).

We explored three different sets of instructional covariates to understand the benefits of combining measures in a composite. First, we examined the composite measure representing students' opportunities to develop analytical, text-based writing skills. Next, we examined each component of the composite individually. Finally, we examined how robust the findings of the composite were to different five-measure combinations of our seven component measures. Our purpose for investigating these composites was to understand whether findings for the composite were sensitive to the inclusion of particular component measures. We describe how robust the results were across 15 different composites representing all of the different potential combinations utilizing five of the seven measures.

Multilevel Multivariate Models

In all of our multivariate models we adjusted for students' prior achievement scale scores in both reading and mathematics. The inclusion of both reading and mathematics prior scores can help reduce bias stemming from measurement error in a single prior achievement score (Rothstein, 2009). We also adjusted for various student background characteristics, including race/ethnicity, gender, free or reduced-price lunch status, and the number of student absences as well as the

students' Individualized Education Plan status (1 = yes student has an Individualized Education Plan). Finally, we also adjusted for a number of teacher and classroom characteristics, such as grade level taught, whether the teacher had obtained a PhD or had advanced certification, and teachers' number of years' experience.

Multivariate Model 1. 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. In this first model, achievement within students was partitioned into two different scales: an RTA achievement scale comprising the five scoring criteria of the RTA and an MSA scale comprising the three MSA Reading subscales—general, literary, and informational. Before running the models, all eight student subscales were standardized to have a mean of zero and standard deviation of 1. 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 et al., 1991). The Level 1 model is described next:

$$\text{ACHIEVE}_{mij} = \psi_{1ij} * (RTA_{mij}) + \psi_{2ij} * (MSA_{mij}) + \varepsilon_{mij} \tag{1.1}$$

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 of the writing rubric; ψ_{1ij} is the average RTA achievement for student i in classroom j; MSA_{mij} is a dummy indicator demarcating the three subscores of the MSA; ψ_{2ij} is the average MSA achievement for student i in classroom j; ε_{mij} is the measurement error for dimension m for student i in classroom j. The Level 2 model is written as follows:

$$\begin{aligned} \psi_{1ij} &= \pi_{10j} + \psi_{1pj} * (A_{pi}) + e_{1ij} \\ \psi_{2ij} &= \pi_{20j} + \pi_{2pj} * (A_{pi}) + e_{2ij} \end{aligned} \tag{1.2}$$

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

$$\begin{aligned} \pi_{10j} &= \beta_{100} + \sum_{q=1}^{4} \beta_{1pq} X_q + \beta_{105}(\textit{Opp. For Analytical Text-Based Writing}_j) + r_{10j} \\ \pi_{20j} &= \beta_{200} + \sum_{q=1}^{4} \beta_{2pq} X_q + \beta_{205}(\textit{Opp. For Analytical Text-Based Writing}_j) + r_{20j} \end{aligned} \tag{1.3}$$

where β_{100} is the average RTA achievement across all classrooms; X_q is a set of (4) teacher and classroom characteristics; β_{1pq} is the association between teacher and classroom characteristics and RTA achievement; 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 association between teacher and classroom characteristics and MSA achievement; r_{20j} is residual error normally distributed with a mean of 0 and a standard deviation of unity.

Our primary focus in these models was the relationship between our measure of students' opportunities-to-learn analytical text-based writing skills and RTA achievement (β_{105}) and between our measure of students' opportunities-to-learn analytical text-based writing skills and MSA achievement (β_{205}) adjusting for student background characteristics including prior reading and math achievement.

Multivariate Model 2. A similar analysis examined a second multivariate model. This model (see Equations 2.1 through 2.3) takes a similar form to the previous multilevel multivariate model with the exception that the scales no longer represent each test separately.

$$\text{ACHIEVE}_{mij} = \psi_{1ij} * ((\text{RTA} + \text{MSA})_{mij}) + \psi_{2ij} * (\text{CONTRAST}_{mij}) + \varepsilon_{mij} \tag{2.1}$$

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

$$\begin{aligned} \pi_{10j} &= \beta_{100} + \sum_{q=1}^{4} \beta_{1pq} X_q + \beta_{105}(\textit{Opp. For Analytical Text-Based Writing}_j) + r_{10j} \\ \pi_{20j} &= \beta_{200} + \sum_{q=1}^{4} \beta_{2pq} X_q + \beta_{205}(\textit{Opp. For Analytical Text-Based Writing}_j) + r_{20j} \end{aligned} \tag{2.3}$$

Instead, the first scale considers achievement on the MSA and RTA together ((RTA + $\text{MSA})_{mij}$). All eight achievement subscores are dummy coded 1 for this scale, and thus ψ_{1ij} is the average achievement across both assessments for student i in classroom j. The second scale (CONTRAST_{mij}) considers the contrast of the two, that is, RTA performance relative to MSA performance. Here, CONTRAST_{mij} is an indicator variable coded $1/n_m$ for each of the five subscores of the RTA and $-1/n_m$ for each of the three 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 MSA for student i in classroom j. Our primary focus in these models were the relationships of our instructional covariates with overall achievement on the RTA and MSA together (β_{105}) and the relationships of our instructional covariates with the contrast between performance on the RTA versus MSA (β_{205}) adjusting for student background characteristics including prior reading and math achievement.

RESULTS

Convergence Among Students' Opportunities to Develop Analytic, Text-Based Writing Skills

Our analysis first examined whether multiple measures provide corroborating evidence regarding students' opportunities to develop analytical text-based writing skills. Results in Table 3 indicate that there is, generally, convergence among the measures. It is interesting that this is especially true among the nonsurvey measures, although even the survey measures were always positively correlated with the other measures (and never below $r = .21$).

Although the degree of overlap of teaching measures via different methods did not allow for a strict multitrait, multimethod comparison, correlations were higher for some cross-method measures, which were a priori hypothesized to have greater overlap. For example, both the surveys and logs captured information about teachers' time devoted to writing. Both methods of data collection also captured aspects of the extent to which teachers integrated reading comprehension and writing (although for the survey measure this was asked in the context of text discussions, and for the logs this captured a more general measure of integration). Correlations between the log and survey measures, broadly capturing similar teaching traits, were generally higher than were correlations between the surveys and assignments where the theoretical overlap was weaker. Likewise, the logs also measured some traits similar with both the cognitive demand and elaborated communication of the assignment measures. Again, these correlations between the log and assignment measures, in general, were higher than were correlations between other measures. As hypothesized in Figure 1, the logs, in general, shared common elements with both the surveys and the assignment measures as demonstrated through the correlation matrix.

TABLE 3
Correlation Between Measures of Students' Opportunity for Extended Writing

	Survey		*Logs*		*Assignments*	
Measure	*Time for Ext. Writing*	*Text Dsc. Emphasis on Writing*	*Log Writing Scale*	*Integration. of Comp. With Writing*	*Task Cognitive Demand*	*% With Open Response*
Survey						
Time for ext. writing	—					
Text dsc. emphasis on writing	.636**	—				
Logs						
Log Writing Scale	.501*	.518*	—			
Integration of comp. with writing	.321	.356	.906**	—		
Assignments						
Task cognitive demand	.284	.446†	.640**	.553*	—	
% with open response	.212	.399	.540*	.377	.669**	—
Length of student written response	.508*	.629**	.665**	.505*	.823***	.738**

Note. Ext. = Extended; Dsc. = Discussion; Comp. = Comprehension.
$^{\dagger}p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$.

TABLE 4
Factor Loadings for Composite of Students' Opportunities to Develop Analytical Text-Based Writing Skills

	Seven-Item Composite		*Six-Item Composite*
	Factor 1	*Factor 2*	*Factor 1*
Opportunities for extended writing composite			
Time for extended writing	.61	.68	—
Text discussion emphasis on writing	.70	.55	.66
Log Writing Scale	.88	−.16	.88
Integration of comprehension with writing	.74	−.34	.75
Task cognitive demand	.83	−.28	.86
% task with open response	.74	−.25	.78
Length of student written response	.91	.02	.91
Eigenvalue	4.24	1.04	3.94
% of variance explained	60.06	14.86	65.69

Factor Structure of Students' Opportunities to Learn Analytic, Text-Based Writing Skills

Although each component of analytic, text-based writing was intended to provide convergent evidence, each component was also measuring different aspects of the opportunity structure hypothesized to be important for students' performance on the RTA. We examined whether these component measures formed a composite with an underlying latent dimension. A principal components analysis to explore the factor structure of the seven component measures revealed two factors with eigenvalues greater than 1 (see Table 4).

Given these results, we decided to exclude the survey item measuring time in extended writing because it had the lowest factor loading. We reran the analysis, which extracted a single factor with an eigenvalue greater than 1[4] and explaining 66% of the variance in the items. Notably, the measure with the lowest loading in the six-item composite was the only remaining survey item. The two component measures with the highest correlations in Table 3 (the Log Writing Scale and the length of student written responses to assignments) had the highest factor loadings. The Cronbach's alpha (.870) indicates high consistency across the items. We proceeded to examine the item composite in prediction models.

Association Between Students' Opportunities to Develop Analytic, Text-Based Writing Skills and Student Learning

Summary of fixed effects in multilevel multivariate Model 1. Our hypotheses stem from the literature demonstrating significant findings for instructional covariates where there is high overlap between what is measured in instruction and the student assessment (D'agostino, Welsh, & Corson, 2007; Leinhardt & Seewald, 1981; Ruiz-Primo, Shavelson, Hamilton, & Klein, 2002). Because students' opportunities to develop analytical text-based writing skills

[4]Examination of the scree plot and the ratio of first to second eigenvalues versus second to third eigenvalues confirms a single factor solution is preferred for the six-item composite.

TABLE 5
Fixed Effects for Multivariate Models with Analytical Text-Based Writing Composite as Covariate

	Multivariate Model 1				*Multivariate Model 2*			
	RTA		*MSA*		*RTA + MSA*		*Contrast RTA/MSA*	
	Coeff	*SE*	*Coeff*	*SE*	*Coeff*	*SE*	*Coeff*	*SE*
Intercept	.024	.056	−.032	.050	.003	.049	.106	.091
Grade	−.094	.107	.183*	.094	.010	.093	−.521**	.158
PhD	.078	.344	.138	.363	.100	.336	−.112	.393
Adv. prof. cert.	.261	.159	−.161	.191	.103	.153	.792*	.299
Years exp.	−.008	.010	−.006	.008	−.007	.009	−.004	.013
OTL Writing	.245***	.069	.049	.066	.171*	.061	.367**	.115
Student Level								
Hispanic	.024	.140	.154	.141	.073	.112	−.245	.328
Black	.117	.162	.040	.158	.088	.131	.145	.362
Native	−.034	.182	−.100	.185	−.059	.148	.125	.417
Asian	.217	.193	.199	.187	.210	.154	.033	.435
Free lunch	−.126*	.062	−.087	.062	−.111*	.051	−.072	.140
Reduced lunch	−.029	.091	.103	.092	.021	.074	−.248	.206
IEP	−.227*	.088	.133	.092	−.092	.072	−.675***	.209
Reading prior ach.	.256***	.040	.459***	.041	.332***	.032	−.381***	.094
Math prior ach.	.191***	.043	.190***	.042	.191***	.034	.003	.098
Absences	−.003	.004	−.012**	.004	−.007*	.003	.016†	.010

Note. RTA = Response-to-Text Assessment; MSA = Maryland School Assessment; Coeff = coefficient; Adv. prof. cert. = advanced professional certification; exp = experience; OTL Writing = Opportunities-to-Learn analytical text- based writing; IEP = Individualized Education Plan; ach. = achievement.

†$p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$.

had significant overlap with the RTA, we were especially interested in the relationship between OTL and student performance on the RTA. As shown in the left half of Table 5, adjusting for prior MSA achievement and student background, students' opportunities to develop analytical text-based writing skills demonstrated a significant relationship with student performance on the RTA ($\beta 105 = .245$, $p = .004$) but not the MSA ($\beta 205 = .049$, $p = .475$). Although grade had an expected effect,[5] no other teacher or classroom variable was significant.

Summary of fixed effects in multilevel multivariate Model 2. We further examined the effects of students' opportunities to develop analytical text-based writing skills in a multivariate model where we examined the relationship between the instructional composite and the RTA and MSA combined (RTA + MSA) and the contrast in performance on the RTA versus the MSA (see right half of Table 5). Students with greater opportunities for analytic, text-based writing had higher combined achievement ($\beta 105 = .171$, $p = .016$) and demonstrated a greater contrast in performance on the RTA versus the MSA ($\beta 205 = .367$, $p = .008$). As predicted,

[5]We expected grade to have an effect on the MSA but not the RTA due to the scale score construction of the MSA and the inability to equate the two different forms of the RTA administered at different grade levels.

TABLE 6
Random Effects for Progression of Multivariate Models with Analytical Text-Based Writing Composite as Covariate

	Null Model	*Student Background*[a]	*Student and Class Background*[a]	*Student/Class Background and Extended Writing*[a]
Multivariate Model 1				
RTA				
b/w students w/in class	.323	.178 (45%)	.178 (45%)	.178 (45%)
b/w class	.176	.105 (40%)	.103 (50%)	.044 (89%)
MSA				
b/w students w/in class	.356	.103 (71%)	.088 (71%)	.103 (71%)
b/w class	.182	.057 (68%)	.035 (81%)	.029 (95%)
Sigma Squared	.477	.477	.477	.477
Multivariate Model 2				
RTA + MSA				
b/w students w/in class	.291	.114 (61%)	.114 (61%)	.114 (61%)
b/w class	.154	.065 (58%)	.056 (64%)	.034 (78%)
Contrast RTA/MSA				
b/w students w/in class	.669	.535 (20%)	.532 (20%)	.533 (20%)
b/w class	.357	.389 (0%)	.179 (50%)	.083 (77%)
Sigma squared	.477	.477	.477	.477

Note. RTA = Response-to-Text Assessment; MSA = Maryland School Assessment.
[a]% Var. explained from null.

amidst a number of insignificant classroom covariates the composite of students' opportunities to learn analytical text-based writing skills was associated with student learning in theoretically relevant ways even after adjusting for student level covariates.

Random effects for multivariate models. To examine the proportion of variance explained in each of our multivariate models we ran a progression of models beginning with a fully unconditional model (null model). The results in Table 6 display the reduction in variance from the null model after entering student background characteristics by themselves, then adding classroom and teacher characteristics, and finally adding students' opportunities to learn analytic, text-based writing skills. The vast majority of between-classroom variance on students' performance on the RTA (89%) and the MSA (95%) is explained by the full model in Multivariate Model 1. However, the proportion of variance explained by background characteristics (50% for the RTA and 81% for the MSA) differs for the two outcomes. After accounting for the variance explained by student and teacher background, students' opportunities to develop analytical, text-based writing by itself explained 57% of the remaining variance between classrooms on the RTA and about 17% of the remaining variance between classrooms on the MSA.[6]

Similar findings were obtained in Multivariate Model 2. About 75% of the between-classroom variance was explained for both the combined measure of achievement and for the contrast in performance on the RTA versus MSA. Students' opportunities to learn analytic, text-based

[6]The percent of variance explained was calculated using the following formula $(\tau_{\beta 0\text{background}} - \tau_{\beta 0\text{ExtendedWriting}})/\tau_{\beta 0\text{background}}$.

writing skills explained 39% of the remaining variance between classrooms in combined achievement and 57% of the remaining variance in the contrast.

Relationship Between Each Individual Component of Students' Opportunities to Develop Analytical Text-Based Writing Skills and Student Learning

We conducted additional exploration of our empirical data to examine how our conclusions would have changed had we only had access to one method of collecting data at a time. Due to issues of cost and burden, researchers and practitioners alike would prefer to collect information on students' opportunities to develop analytical, text-based writing skills as efficiently as possible. The following results allowed us to compare and contrast results from individual components with the composite measure examined previously.

Summary of findings for Multivariate Model 1. For ease of comparison, in Table 7 we report only the results for individual components even though all models contain the same student and classroom characteristics as the previous models. We focus our comparison of individual component measures on columns containing the p value and the percent of variance explained.

Our results indicate positive (often nonsignificant) relationships between most of our component measures and student learning outcomes. Results from the left half of Table 7 reveal marginally significant relationships ($p < .10$) with the RTA for three of the seven measures (Log Writing Scale, cognitive demand of the assignment, and proportion of assignments allowing an extended open response) and a significant finding for just one of the measures—length of students' written responses to assignments ($\beta 105 = .261$, $p = .002$). For these four measures, the amount of additional variance explained in student performance on the RTA ranges from 21% to 57%. Neither of the components of students' OTL derived from the annual survey was predictive of student achievement. Moreover, no single measure is associated with achievement on the MSA after adjusting for student background including prior achievement.

Summary of findings for Multivariate Model 2. Results from the second multivariate model (displayed in the right half of Table 7) reveal an expected pattern given the results from the first multivariate model. Only one of the seven covariates reached a level of significance ($p < .05$) on combined achievement (RTA + MSA), and only two achieved significance on the contrast between student performance on the RTA relative to their performance on the MSA. It is interesting to note that both of these measures were from the collection of assignments[7]—cognitive demand of the assignment was related to the contrast only ($\beta 205 = .278$, $p = .048$), whereas the length of students' written responses was related to both combined achievement ($\beta 205 = .194$, $p = .006$) and the contrast ($\beta 205 = .334$, $p = .020$).

Comparing and contrasting composite of students' opportunities to develop analytical, text-based writing skills and individual components. Although a similar general pattern of effects holds for the individual components as it does for the composite, the inferences

[7]The assignments, in particular, should be most aligned to the RTA because we asked teachers to turn in challenging assignments and because this represented students' opportunities to practice the skills required to do well on the RTA.

TABLE 7
Relationship Between Individual Components of Analytical Text-Based Writing and Achievement on RTA and MSA

	Multivariate Model 1							
	RTA (β_{105})				*MSA (β_{205})*			
	Coeff	*SE*	*p*	*% Var Exp*[a]	*Coeff*	*SE*	*p*	*% Var Exp*[a]
Time for ext writing	.020	.115	.863	1.1	.054	.083	.526	6.1
Text dsc emphasis on writing	.177	.116	.154	15.2	.011	.088	.901	1.0
Log Writing Scale	.191	.098	.075†	21.2	.049	.074	.522	3.4
Intgrtn of Comp with Writing	.107	.085	.232	10.2	.037	.060	.550	2.9
Task cog demand	.154	.079	.074†	22.1	.006	.069	.932	1.2
% task w/open response	.158	.074	.053†	25.2	.037	.067	.595	6.2
Length of student written response	.261	.065	.002**	57.6	.083	.070	.261	17.1
	Multivariate Model 2							
	RTA + MSA (β_{105})				*Contrast RTA/MSA (β_{205})*			
Time for ext writing	.033	.095	.733	2.4	−.063	.172	.720	0.2
Text dsc emphasis on writing	.115	.095	.251	10.1	.311	.186	.121	24.0
Log Writing Scale	.137	.080	.111	17.1	.265	.158	.119	21.1
Integration of Comp With Writing	.081	.068	.259	9.0	.131	.135	.351	8.0
Task cognitive demand	.099	.068	.173	14.5	.278	.126	.048*	35.3
% Task w/open response	.112	.063	.102	21.0	.226	.127	.101	24.3
Length of student written response	.194	.059	.006**	50.8	.334	.125	.020*	45.6

Note. RTA = Response-to-Text Assessment; MSA = Maryland School Assessment.

[a]Percent variance explained is calculated from model immediately prior to the inclusion of the single covariate measuring instruction and is thus calculated from variance remaining after adjusting for all student and teacher characteristics.

$^{\dagger}p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$.

we can draw from each set of results differs. Individual components were variable in their relationship to student achievement on the RTA. In this sample, only if you chose the component measuring the length of students' written response would you infer a significant association between students' opportunities for elaborated communication and learning. However, the findings from our composite measure demonstrate that students with greater opportunities to develop analytical text-based writing skills have significantly higher achievement on the RTA, on combined achievement on the RTA plus MSA, and they score significantly higher on the RTA relative to their performance on the MSA.

Tests for Robustness of Findings: Different Composite Configurations

Finally, to ensure our findings are not attributable to the particular scale construction or to one particular covariate, we examined whether our findings were robust to different combinations

of our component measures of analytical text-based writing. As shown in Table 8, no matter the configuration of five individual component measures, the resulting composites all had a consistent factor structure. As predicted given the correlation matrix (see Table 3), each of 15 different combinations have moderate to high Cronbach's alpha ranging from a low of .74 for combination 14 to a high of .87 for combination 7. The percentage of variance in the items explained by the first factor using principal components extraction and forcing a single factor ranged from a low of 50% to a high of 67%.

We next examined each of these combinations as independent variables in our multivariate models. The results were robust to the different combinations and similar to those reported from the six-item composite of students' opportunities to develop proficiency at producing analytical, text-based writing. In each case, the combination of measures produced at least marginally significant effects on the RTA and no significant findings on the MSA. Furthermore, in each case the factor was predictive of overall achievement when the RTA and MSA were combined. Finally, in all but one case, the factors were at least marginally significant on the contrast between the RTA and MSA.

What is surprising about these findings is the consistency of the results, although it should be noted that each composite has at least one component from each of the assignments, logs, and surveys, and thus each method of data collection was always represented on each composite.

DISCUSSION

Our work was motivated by the desire to both demonstrate and describe how students' opportunities to learn in upper elementary language arts classrooms contribute to their learning. This work connects to two larger purposes for understanding the consequences of students' opportunities to learn. First, description of the literacy practices that are associated positively with student learning provides an empirical basis for debating and deciding on important elements for student curricula. But this is also tightly entwined with our goals for teaching practice because teaching is a primary contributor to student learning (Nye et al., 2004; Rivkin et al., 2005; Sanders et al., 1997). Thus, identifying and describing aspects of teaching related to student learning also forms the basis for defining a professional learning agenda involving both teacher education and designs for improving teaching practice. Second, the concept of opportunity to learn also carries notions of fairness and equity. Broadening our understanding of how students' opportunities in early literacy shape their learning in specific ways could inspire policymakers to both monitor and intervene on the current opportunity structure.

Our work can be seen as both constraining the measurement of students' opportunities to learn and at the same time broadening it. On one hand, we anchored our measures of teaching practice around a specific, but complex, student skill. Because writing in response to text is a foundational skill for later school success, and because it is aligned with the CCSS, we explored students' ability to form an analytical, text-based argument and to communicate their argument in writing. Because we sought to measure teaching practices associated with developing this skill in students, we created our own project-based response-to-text assessment. Hence, we constrained the domain of literacy instruction in the sense that, for the analyses presented here, we sampled teaching practice related only to this specific skill.

TABLE 8
Relationship Between 15 Different Combinations of Analytical Text-Based Writing Composites and Achievement on the RTA and MSA

			Multivariate Model 1							
			RTA (β_{105})				*MSA (β_{205})*			
	% Var Exp	*α*	*Coeff*	*SE*	*p*	*% Var Exp*[a]	*Coeff*	*SE*	*p*	*% Var Exp*[a]
Composites[b] with factor analytic information										
1	58%	.81	.189	.091	.062†	23.8	.044	.070	.537	3.0
2	62%	.84	.236	.091	.023*	33.4	.076	.074	.325	8.7
3	55%	.77	.206	.093	.047*	26.4	.063	.074	.410	6.2
4	56%	.78	.188	.082	.040*	27.4	.047	.065	.482	4.4
5	63%	.84	.215	.081	.022*	33.6	.060	.067	.390	6.6
6	60%	.82	.231	.082	.016*	37.6	.076	.071	.304	10.9
7	67%	.87	.221	.079	.016*	36.6	.050	.069	.483	4.6
8	60%	.83	.198	.079	.028*	30.9	.038	.066	.580	2.9
9	64%	.85	.238	.079	.011*	41.0	.065	.072	.383	8.3
10	61%	.83	.247	.083	.012*	39.2	.060	.074	.430	6.0
11	53%	.77	.225	.085	.022*	33.6	.046	.073	.537	3.7
12	57%	.81	.269	.086	.009**	44.7	.081	.080	.327	11.1
13	58%	.81	.223	.082	.018*	35.2	.056	.070	.433	5.7
14	50%	.74	.200	.083	.032*	29.6	.042	.068	.545	3.4
15	54%	.78	.246	.083	.062†	40.8	.077	.075	.320	10.9
			Multivariate Model 2							
			RTA + MSA (β_{105})				*Contrast RTA/MSA (β_{205})*			
Composites[b] with factor analytic information										
1	58%	.81	.134	.074	.096†	18.7	.270	.150	.097†	25.1
2	62%	.84	.176	.075	.037*	29.0	.300	.153	.074†	27.6
3	55%	.77	.153	.077	.070†	22.4	.269	.153	.104	23.0
4	56%	.78	.135	.067	.068†	22.2	.264	.136	.076†	27.3
5	63%	.84	.157	.067	.038*	27.9	.290	.139	.059†	31.1
6	60%	.82	.173	.069	.029*	33.0	.290	.139	.059†	29.8
7	67%	.87	.157	.066	.036*	28.7	.322	.135	.035*	39.2
8	60%	.83	.138	.066	.060†	23.3	.300	.133	.043*	35.9
9	64%	.85	.173	.068	.026*	34.0	.324	.133	.031*	38.3
10	61%	.83	.177	.071	.027*	31.4	.350	.146	.034*	39.8
11	53%	.77	.158	.072	.047*	25.8	.336	.146	.040*	37.7
12	57%	.81	.199	.075	.021*	38.3	.352	.153	.030*	38.5
13	58%	.81	.161	.068	.036*	28.4	.313	.143	.049*	35.0
14	50%	.74	.141	.068	.061†	21.8	.296	.142	.058†	32.8
15	54%	.78	.183	.072	.025*	35.4	.316	.140	.044*	33.9

Note. RTA = Response-to-Text Assessment; MSA = Maryland School Assessment.

[a]Percent variance explained is calculated from model immediately prior to the inclusion of the single covariate measuring instruction and is thus calculated from variance remaining after adjusting for all student and teacher characteristics. [b]Composites were made up of the following individual covariates: (a) Time for extended writing, (b) Text discussions with emphasis on writing, (c) Log writing scale, (d) Integration of comprehension with Writing, (e) Task cognitive demand, (f) Percentage of tasks with open response, and (g) Length of students' written responses, such that 1 = abcde; 2 = abcdg; 3 = abcdf; 4 = acdeg; 5 = acdef; 6 = acdfg; 7 = bcdeg; 8 = bcdef; 9 = bcdfg; 10 = abceg; 11 = abcef; 12 = abcfg; 13 = abdeg; 14 = abdef; 15 = abdfg.

†$p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$.

At the same time, we broadened our notion of the opportunity structure when we sampled within this domain. Students' performance on the RTA was hypothesized in Figure 1 to be the product of accumulated learning opportunities in writing, including the opportunity to engage in high cognitive demand tasks allowing for elaborated communication, the opportunity to engage in analytic comprehension and more time spent engaged in all forms of writing. Because our conception of the opportunity structure included both performance-based elements measuring the quality of the tasks students were provided as well as measures of time on content (i.e., how tasks were embedded in literacy instruction throughout the year), we used surveys, instructional logs, and teacher assignments with student work to capture teaching practice in writing. Our measures represent a broader conception of the opportunity structure, but they also represent various methods and formats for collecting such data.

Both our conceptualization of students' opportunities to learn analytic, text-based writing and the methods used to study it represent potential reasons why multiple measures may be advantageous for researchers and practitioners to consider going forward. On one hand, multiple methods could contribute to learning about sources of measurement error in data collection efforts and could also help reduce measurement error when composites are generated. Triangulating measures from different sources could help identify whether measurement bias is contained in any particular measure and contribute to our understanding of the interrelationships of the measures. Although triangulation is often assumed to assess convergence of measures, it is equally important to document inconsistent as well as contradictory information (Mathison, 1988). Researcher explanations for observed evidence, whether convergent or divergent, hold value for their potential to advance theory about how teaching could be accurately measured with fidelity to its complexity and ways in which teaching is associated with student learning.

At the same time, multiple measures seem important to protect against construct-under-representation. Because teaching is a complex act, it is important to assess it in its complexity. Consider, for example, parallels to the use of performance assessments of student learning where complex educational goals are the focus of measurement. Here, "authentic" tasks have inherent strengths because they represent direct measures of the target domain and "the directness of the interpretation makes many competing interpretations . . . seem implausible" (Kane, Crooks, & Cohen, 1999, p. 5). As a result, it is easy to extrapolate scoring inferences to the target domain to contribute to an interpretive argument for the validity of inferences based on score interpretations. An interpretive argument, however, is only as strong as the weakest link in the chain of evidence from observation to observed score, from observed score to universe score, and from universe score to target score (Kane et al., 1999). Evidence for validity is strongest when available evidence allows for strong generalizations while maintaining the ability to extrapolate to the target domain. Thus, although it is important to sample as directly as possible from the target domain, difficulties arise when attempting to rate or score complex assessments. The greatest difficulty, therefore, is in generating evidence to support generalizations from the observed score to the universe score, that is, to ensure that scores of complex activities are representative of the target domain.

The prior discussion reviews prior work on performance assessments for student learning but at the same time summarizes current tensions in the field when considering how best to measure teaching. For example, observations/videos of teaching offer high potential for extrapolation to the target domain, especially when observers can attend to teaching in its complexity and then accurately quantify it. However, observations are also costly and questions remain about

the extent to which they can create generalizations from the observed score to the universe score. This tension results both because it is difficult to observe multiple occasions and because some aspects of instruction, and perhaps those researchers are most interested in (in this case writing), occur infrequently and as a result may never be observed.

We have presented an alternative approach to measuring writing instruction in classrooms by collecting and triangulating components of students' opportunity to develop text-based writing from assignment artifacts, daily logs, and surveys. Our choice of data collection was motivated as much by issues of researcher burden as it was by the ability to generalize from the observed score to the universe score in order to bolster our validity argument. In the remainder of the article we consider both how these decisions contributed to our findings and whether and how our experience combining measures can inform the measurement of teaching.

Review of Findings

Our findings demonstrated convergence among our measures of students' opportunities to develop analytic, text-based writing, resulting in the formation of a composite. Triangulation across assignment tasks, logs, and surveys revealed a few insights. Descriptive statistics of the individual component measures revealed students' opportunities to develop analytic, text-based writing were infrequent compared to other strategies taught in language arts and, on average, tasks only required recall of text. The pattern of correlations revealed modest correlations across all measures, with sensible findings including higher correlations between the log measures with both the surveys and tasks than between the surveys and tasks, confirming a priori theory about our measures. Finally, in general, correlations were weakest for the survey measures, suggesting that annual self-reports may not be the optimal way to measure differences across classrooms in writing.

Prediction models confirmed an association between our composite measuring students' opportunities to develop analytic, text-based writing and performance on the RTA (but not the MSA by itself). Moreover, students' with greater opportunities to develop analytic, text-based writing demonstrated higher performance on the RTA plus MSA and demonstrated higher performance on the RTA relative to performance on the MSA.

We note several observations with respect to these findings. First, the composite measure demonstrated a significant association between teaching and learning; however, all but one of the individual measures by themselves failed to do so under traditional levels of statistical significance ($p < .05$). Thus, combining measures to represent the broader opportunity structure was important in this case for demonstrating a teaching–learning association. Second, these findings depend on the measurement of classroom performance on the RTA. Alignment between measures of teaching and learning are paramount for demonstrating effects on teaching. Third, when we examined measures individually, neither survey measure predicted student learning, only one of the log measures was marginally significant ($p < .10$), two of the assignment measures were marginally significant, and the third assignment measure (average length of student response) was statistically significant. The assignment measures, therefore, seemed to carry relatively stronger signal than the log measures, and both the log and assignment measures carried stronger signal than the survey measures. Fourth, although surveys represent the most feasible method for collecting data on teaching practice, low intercorrelations among other measures of teaching practice and the lack of signal in prediction models demonstrate distinct

advantages for more in-depth measures of practice such as those collected through tasks and instructional logs.

Combining multiple measures provided several advantages in our analyses. First, they helped corroborate information and at the same time broadened construct representation to approximate the target domain. The measures chosen were more suited to measuring writing than other alternatives such as observations. In addition, prediction models of the various combinations of individual covariates demonstrated significant findings that were robust to the different potential combinations. When various methods were represented in composites the results were remarkably consistent. Furthermore, findings that the combined measures explained so much of the remaining between classroom variance after accounting for prior achievement and student background indicate that combined measures left less unexplained error in the achievement outcomes.

Reflections about our experience combining multiple measures lead to two insights. First, optimal methods for measuring teaching likely interact with the focal aspect of what is being measured in instruction. In our case, measuring analytic, text-based writing instruction was successfully captured through artifacts and daily logs, whereas elsewhere videos have been used to measure reading comprehension instruction (Carlisle et al. 2011). Researchers' choice of method should be sensitive to the context of what is being measured. Second, researchers will need to strike the balance between intended convergence of multiple measures and construct representation from a target domain. Although in our case we achieved convergence while measuring a broad opportunity structure for a complex skill, this will not always be the intended goal. Regardless of whether convergence is achieved, multiple measures will be important to further develop theories of teaching and to further investigate how teaching relates to student learning.

Finally, we cannot ignore the fact that multiple *learning* measures were instrumental for demonstrating the influence of teaching on students (Resnick & Resnick, 1992). The MSA by itself was not sensitive to our measure of students' opportunities to develop analytical text-based writing skills, but the project-developed RTA was. State accountability tests such as the MSA cannot evaluate all of the knowledge and skills students need to be academically successful. Moreover, research shows that many state accountability tests are more successful at measuring lower level skills than higher level skills (Rothman, Slattery, Vranek, & Resnick, 2002). In the context of CCSS implementation, researchers are likely to continue to focus on teaching behaviors that are designed to develop students' higher level thinking skills, the teaching skills that often are the most difficult to enact. It is important to bear in mind, however, that students' scores on state accountability tests will not necessarily be sensitive to these teaching practices.

Limitations

The small sample size reveals several limitations to our data and findings. First and foremost, the small sample size translates into limited power to detect all but very large relationships in our data (i.e., relationships with effect sizes greater than about .5 at power $1 - \beta = .8$). Given the lack of power, it is encouraging that we found evidence of a strong association between the composite measure of students' opportunities to develop analytical, text-based writing skills and student performance on the RTA. However, it is not yet clear whether the measures of teaching will predict classroom performance on the MSA in subsequent years with larger numbers of

classrooms. Replication studies will allow for further investigations into how different measures of teaching differentially predict multiple student learning outcomes. They will also seek to learn if the association between students' opportunities for analytical text-based writing skills and the RTA generalize to different samples of students and teachers.

Furthermore, in larger samples, exploratory and confirmatory factor analyses could be used to examine the covariance structure of different measures of teaching practice. Here we confined our analysis to a single composite measure of the opportunities students had to develop analytical, text-based writing skills. In the future we might simultaneously consider the frequency and intensity of students' opportunities to learn how to comprehend text, incorporating data from multiple methods. Using multilevel multivariate models, second-order factor models, or bifactor models with larger samples, it will be possible to explore multidimensional aspects of teaching. Those methods would produce covariates that could then be used to understand canonical correlations with multidimensional measures of student learning.

CONCLUSION

This work has several implications for the field as we embark on further attempts to identify teaching–learning associations. First, it is wise to consider both sides of the teaching–learning connection. Having a measure of student learning that captures a specific element of literacy learning may be vital for demonstrating how teaching practice manifests differences in student achievement across classrooms. It may be worth investing in measures of student learning that go beyond the convenient and easy-to-collect state accountability tests because alternative measures, such as the RTA demonstrated here, are apt to be more sensitive to important between-classroom differences in the teaching opportunity structure.

Second, we chose to focus on an aspect of literacy instruction and learning that would be considered higher order because providing "high" literacy for all (Bereiter & Scardamalia, 1987; Resnick, 1987, 2010) is part of the American ethos. It remains to be seen how much our findings were due to the fact we examined teaching and learning that resides outside of the current accountability framework. As a result, students were far from advanced on the assessed skill and teaching practice was also highly variable. Research such as this should raise the question, therefore, of whether students' ability to form an analytical argument in response to text, and by extension teaching practice known to align with students' ability on that skill, should be part of an accountability framework. Although we think incorporating analytical, text-based writing as a goal of language arts teaching and learning would advance students' learning, in general, the larger point is that only by identifying and pairing specific teaching–learning associations can we begin to debate the relative merits of different curricula and begin aligning efforts at improvement toward specific teaching and learning goals.

Finally, this work also has implications for the measurement of teaching because we were able to successfully combine multiple measures to represent a fairly complex opportunity structure within the domain of language arts instruction. We see this as keeping in the spirit of "getting it right" fostered by Leigh Burstein, a pioneer in the opportunity to learn literature (Shavelson & Webb, 1995). In this way we resisted the temptation to simplify by confining our measure of teaching to only "content coverage" and instead tried to capture the complexity of teaching (combining content coverage with both instructional methods and ratings of quality) as

faithfully as we could (Shavelson & Webb, 1995). We invite others to extend this work further as we work toward measuring the complexity of instruction and understanding its relations with measures of student learning.

ACKNOWLEDGMENTS

This research was supported in part by grants from the W.T. Grant Foundation and the Spencer Foundation. The opinions expressed in the article are those of the authors, not the sponsors. The authors remain responsible for any errors in the work.

REFERENCES

American Institute for Research. (2005). *Rigor, relevance and results: The quality of teacher assignments and student work in new and conventional high schools*. Washington, DC: American Institutes for Research and SRI International.

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

Baker, E. (2007). The end(s) of testing. *Educational Researcher, 36*, 309–317.

Benko, S. L. (2012). *Teaching to the task: Preservice teachers' instruction for cognitively demanding writing tasks* (Unpublished doctoral dissertation). Pittsburgh, PA: University of Pittsburgh.

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

Boscolo, P., & Carotti, L. (2003). Does writing contribute to improving high school students' approach to literature? *Educational Studies in Language and Literature*, *3*, 197–224.

Brewer, D. J., & Stacz, C. (1996). *Enhancing opportunity to learn measures in NCES data.* Santa Monica, CA: RAND.

Brophy, J., & Good, T. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), *Handbook of research on teaching, 3rd ed*. (pp. 328–375). New York, NY: Macmillan.

Calkins, L. (1986). *The art of teaching writing.* Portsmouth, NH: Heinemann.

Calkins, L. Ehrenworth, M., & Lehmen, C. (2012). *Pathways to the common core: Accelerating achievement*. Portsmouth, NH: Heinemann.

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*, 409–439.

Carroll, J. (1963). A model of school learning. *Teachers College Record*, *64*, 723–733.

Common Core State Standards Initiative. (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

Correnti, R. (2005). *Literacy instruction in CSR schools: Consequences of design specification on teacher practice* (Unpublished doctoral dissertation). University of Michigan, Ann Arbor.

Correnti, R., Phelps, G., & Kisa, Z. (2010). *Investigating the relationship between teachers' knowledge, literacy practice and growth in student learning*. Paper presented at 2010 annual meeting of the American Educational Research Association, Denver, CO.

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.

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. *The 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.

Doyle, W. (1983). Academic work. *Review of Educational Research, 53*, 159–199.

Dunkin, M., & Biddle, B. (1974). *The study of teaching*. New York, NY: Holt, Reinhart and Winston.

Gamoran, A., Porter, A., Smithson, J., & White, P. (1997). Upgrading high school mathematics instruction: Improving learning opportunities for low-achieving, low-income youth. *Educational Evaluation and Policy Analysis*, *19*, 325–338.

Graham, S., & Harris, K. R. (1993). Self-regulated strategy development: Helping students with learning problems develop as writers. *Elementary School Journal, 94*, 169–181.

Graham, S., & Perin, D. (2007). *Writing next: Effective strategies to improve writing of adolescents in middle and high schools—A report to Carnegie Corporation of New York.* Washington, DC: Alliance for Excellent Education.

Grossman, P. L., Stodolsky, S. S., & Knapp, M. (2004). *Making subject matter part of the equation: The intersection of policy and content* (Occasional Paper, Document 0-04-1). Center for the Study of Teaching and Policy, University of Washington, Seattle.

Herman, J., Klein, D., & Abedi, J. (2000). Assessing students' opportunity to learn: Teacher and student perspectives. *Educational Measurement: Issues and Practice, 19*(4), 16–24.

Hill, H. C., Rowan, B., & Ball, D. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal*, *42*, 371–406.

Hillocks, G., Jr. (1984). What works in teaching composition: A meta-analysis of experimental treatment studies. *American Journal of Education, 93*(1), 107–132.

Hillocks, G., Jr. (2006). Middle and high school composition. In P. Smagorinsky (Ed.), *Research on composition: Multiple perspectives on two decades of change* (pp. 48–77). New York, NY: Teachers College Press.

Husen, T. (1974). Multi-national evaluation of school systems: Purposes, methodology, and some preliminary findings. *Scandinavian Journal of Educational Research*, *18*(1), 13–39.

Jennings, J. F. (1998). *Why national standards and tests? Politics and the quest for better schools*. Thousand Oaks, CA: Sage.

Junker, B. W., Matsumura, L. C., Crosson, A., Wolf, M. K., Levison, A., Wiesberg, J., & Resnick, L. (2006). *Overview of the Instructional Quality Assessment* (CSE Tech. Rep. 671). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, *18*(2), 5–17.

Klein, P. (1999). Reopening inquiry into cognitive processes in writing-to-learn. *Educational Psychology Review*, *11*, 203–270.

Knapp, M., Adelman, N., Marder, C., McCollum, H., Needels, M., Padilla, C., ... Zucker, A. (1995). *Teaching for meaning in high poverty classrooms.* New York, NY: Teachers College Press.

Langer, J. A., & Applebee, A. (1987). *How writing shapes thinking: A study of teaching and learning* (Research Rep. No. 22). Urbana, IL: National Council on Teachers of English.

Lee, J., Grigg, W., & Donahue, P. (2007). *The Nation's Report Card: Reading 2007*. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

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

Marshall, J. (1987). The effects of writing on students' understanding of literary texts. *Research in the Teaching of English, 21*, 30–63.

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

Mathison, S. (1988). Why triangulate? *Educational Researcher*, *17*(2), 13–17.

Matsumura, L. C., Garnier, H., Pascal, J., & Valdés, R. (2002). Measuring instructional quality in accountability systems: Classroom assignments and student achievement. *Educational Assessment*, *8*, 207–229.

Matsumura, L. C., Garnier, H. E., Slater, S. C., & Boston, M. D. (2008). Toward measuring instructional interactions "at-scale." *Educational Assessment, 13*, 267–300.

Matsumura, L. C., Patthey-Chavez, G., Valdes, R., & Garnier, H. (2002). Teacher feedback, writing assignment quality, and third-grade students' revision in lower- and higher-achieving urban schools. *The Elementary School Journal, 103*, 3–25.

Matsumura, L. C., Slater, S. C., & Crosson, A. (2008). Classroom climate, rigorous instruction and curricula, and students' interactions in urban middle school classrooms. *Elementary School Journal*, *103*, 293–312.

McKnight, C., Crosswhite, F., Dossey, J., Kifer, E., Swafford, J., Travers, K., & Cooney, T. (1987). *The underachieving curriculum: Assessing U.S. school mathematics from an international perspective*. Champaign, IL: Stipes.

Medley, D. (1977). *Teacher competence and teacher effectiveness: A review of process-product research*. Washington, DC: American Association of Colleges for Teacher Education.

Morrow, L. (1992). The impact of a literature-based program on literacy achievement, use of literature, and attitudes of children from minority backgrounds. *Reading Research Quarterly, 27*, 251–275.

Muthén, L. K., & Muthén, B. O. (1998–2010). *Mplus User's guide. Sixth Edition*. Los Angeles, CA: Author.
National Council on Education Standards and Testing. (1992). *Raising standards for American education.* Washington, DC: U.S. Government Printing Office.
Newmann, F., Bryk, A., & Nagaoka, J. (2001). *Authentic intellectual work and standardized tests: Conflict or coexistence?* Chicago, IL: Consortium on Chicago School Research.
Newmann, F. M., Lopez, G., & Bryk, A. S. (1998). *The quality of intellectual work in Chicago schools: A baseline report*. Chicago, IL: Consortium on Chicago School Research.
Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*, 237–257.
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.
Pressley, M., Allington, R. L., Wharton-McDonald, R., Block, L. C., & Morrow, L. (2001). *Learning to read: Lessons from exemplary first-grade classrooms.* New York: Guilford.
Porter, A. (1993). School delivery standards. *Educational Researcher, 22*(5), 24–30.
Porter, A. (1995). The uses and misuses of opportunity-to-learn standards. *Educational Researcher, 24*(1), 21–27.
Raudenbush, S. W., & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.
Raudenbush, S. W., Rowan, B., & Kang, S. J. (1991). A multilevel, multivariate model for studying school climate with estimation via the EM algorithm and application to US 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., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement, and instruction* (pp. 37–75). Boston, MA: Kluwer.
Rivkin, S. G., Hanushek, E., & Kain, J. F. (2005). Teachers, schools and academic achievement. *Econometrica*, *73*, 417–458.
Rothman, R. (2011). *Something in common: The Common Core standards and the next chapter in American education.* Cambridge, MA: Harvard Education.
Rothman, R., Slattery, J. B., Vranek, J. L., & Resnick, L. B. (2002). *Benchmarking and alignment of standards and testing* (CSE Tech. Rep. 566). Los Angeles: Center for the Study of Evaluation, University of California, Los Angeles.
Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy, 4*, 537–571.
Rowan, B., Correnti, C., & Miller, R. J. (2002). What large-scale, survey research tells us about teacher effects on student achievement: Insights from the *Prospects* study of elementary schools. *Teachers College Record, 104*(8), 1525–1567.
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.
Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student academic achievement*. Knoxville: University of Tennessee Value-Added Research and Assessment Center.
Sanders, W. L., Wright, S. P., & Horn, S. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education*, *11*(1), 57–67.
Schmidt, W., & McKnight, C. (1995). Surveying educational opportunity in mathematics and science: An international perspective. *Educational Evaluation and Policy Analysis, 17*, 337–353.
Shanahan, T., & Shanahan, C. (2008). Teaching disciplinary literacy to adolescents: Rethinking content-area literacy. *Harvard Educational Review*, *78*(1), 40–59.
Shavelson, R., & Webb, N. (1995). On getting it right. *Educational Evaluation and Policy Analysis, 17*, 275–279.
Stodolsky, S. S., & Grossman, P. A. (1995). The impact of subject matter on curricular activity: An analysis of five academic subjects. *American Educational Research Journal*, *32*, 227–249.
Tierney, R., Soter, J., O'Flahavan, J., & McGinley, W. (1989). The effects of reading and writing upon thinking critically. *Reading Research Quarterly, 24*, 134–173.