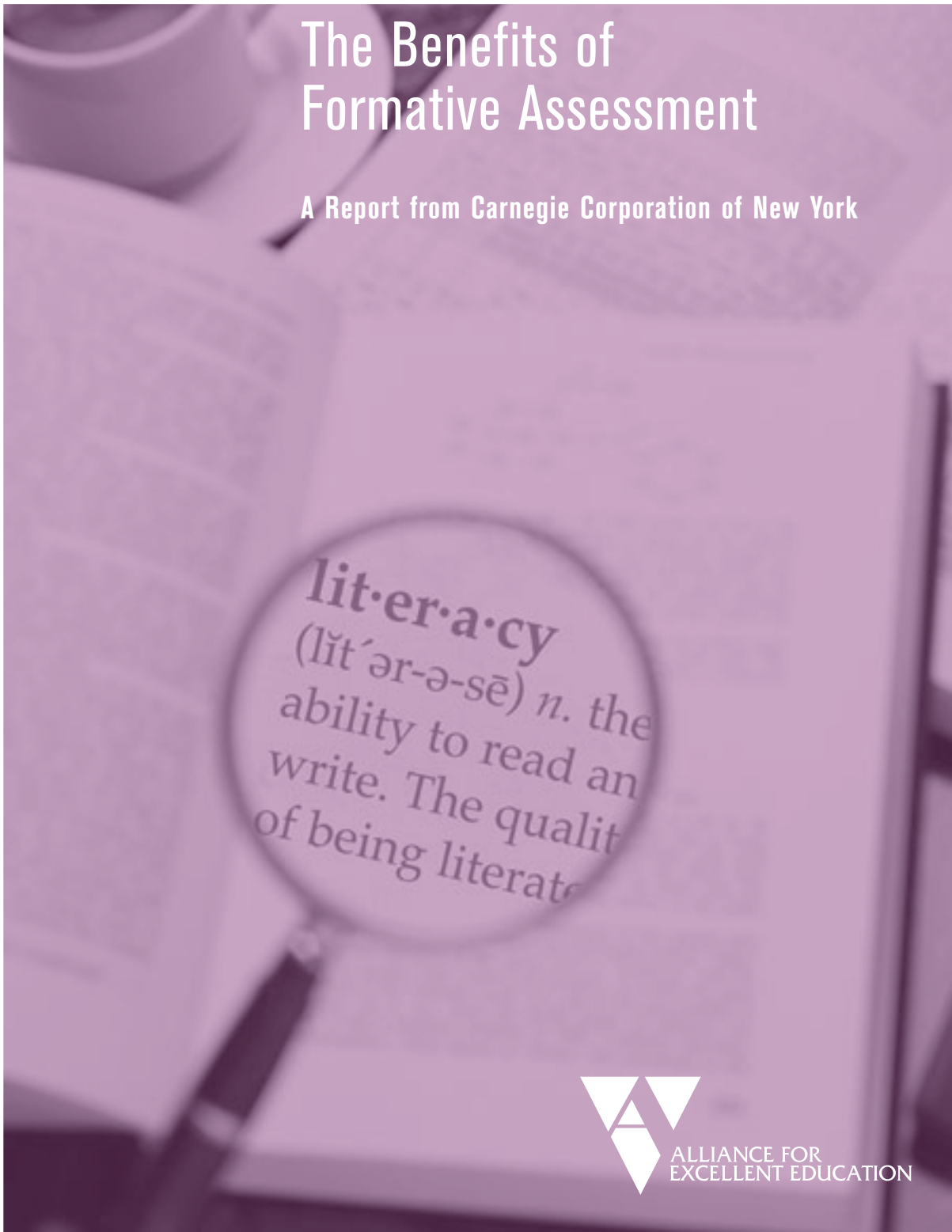
Informing Writing

The Benefits of Formative Assessment

A Report from Carnegie Corporation of New York



Steve Graham
Karen Harris
Michael Hebert

Vanderbilt University

ALLIANCE FOR EXCELLENT EDUCATION

Carnegie Corporation's *Advancing Literacy* program is dedicated to the issues of adolescent literacy and the research, policy, and practice that focus on the reading and writing competencies of middle and high school students. *Advancing Literacy* reports and the *Time to Act* series of publications are designed to encourage local and national discussions, explore promising ideas, and incubate models of practice, but do not necessarily represent the recommendations of the Corporation. For more information, visit www.carnegie.org/literacy.

Published by the Alliance for Excellent Education.

Suggested citation: Graham, S., Harris, K., and Hebert, M. A. (2011). *Informing writing: The benefits of formative assessment. A Carnegie Corporation Time to Act report*. Washington, DC: Alliance for Excellent Education.

Informing Writing

The Benefits of Formative Assessment

Steve Graham, Karen Harris, and Michael Hebert
Vanderbilt University

A Report from Carnegie Corporation of New York

About the Alliance for Excellent Education

The Alliance for Excellent Education is a Washington, DC–based national policy and advocacy organization that works to improve national and federal policy so that all students can achieve at high academic levels and graduate high school ready for success in college, work, and citizenship in the twenty-first century. The Alliance focuses on America's six million most-at-risk secondary school students—those in the lowest achievement quartile—who are most likely to leave school without a diploma or to graduate unprepared for a productive future.

The Alliance's audience includes parents, educators, the federal, state, and local policy communities, education organizations, business leaders, the media, and a concerned public. To inform the national debate about education policies and options, the Alliance produces reports and other materials, makes presentations at meetings and conferences, briefs policymakers and the press, and provides timely information to a wide audience via its biweekly newsletter and regularly updated website, www.all4ed.org.

About Carnegie Corporation of New York

Carnegie Corporation of New York, which was established by Andrew Carnegie in 1911 "to promote the advancement and diffusion of knowledge and understanding," is one of the oldest, largest, and most influential of American grantmaking foundations. The foundation makes grants to promote international peace and to advance education and knowledge—primary concerns to which founder Andrew Carnegie devoted the foundation. For more information, visit www.carnegie.org.

The Authors

Steve Graham, EdD, is the Currey Ingram Professor of Special Education and Literacy, a chair he shares with Karen R. Harris, at Vanderbilt University's Peabody College of Education. Dr. Graham's research interests include how writing develops, why some students have difficulty mastering this critical skill, and the identification of effective writing practices. He is the former editor of *Exceptional Children* and *Contemporary Educational Psychology* and has written more than 250 publications, including *Handbook of Writing Research, Best Practices in Writing Instruction, Writing Better, Powerful Writing Strategies for All Students, Handbook of Learning Disabilities,* and the *APA Handbook of Educational Psychology* (in press). He is the coauthor of the influential meta-analyses of writing interventions *Writing Next* and *Writing to Read*, funded by Carnegie Corporation of New York. Dr. Graham is the recipient of the Council of Exceptional Children's Career Research Award, the Samuel A. Kirk Award from the Division of Learning Disabilities, and the Distinguished Research Award from the Special Education Interest Group of the American Educational Research Association.

Karen R. Harris, EdD, is the Currey Ingram Professor of Special Education and Literacy, a chair she shares with Steve Graham, at Vanderbilt University's Peabody College of Education. Dr. Harris's research interests include writing development, self-regulation, and the identification of effective writing practices for typically developing and struggling writers. She is the former editor of the *Journal of Educational Psychology* and has written more than 200 publications, including *Handbook of Learning Disabilities, Writing Better, Powerful Writing Strategies for All Students,* and the *APA Handbook of Educational Psychology* (in press). She is the developer of the Self-Regulated Strategy Development model, which has been tested in over fifty studies with a wide range of writers. Dr. Harris is the recipient of the Council of Exceptional Children's Career Research Award, the Samuel A. Kirk Award from the Division of Learning Disabilities, and the Distinguished Research Award from the Special Education Interest Group of the American Educational Research Association.

Michael Hebert is a doctoral student in special education at Vanderbilt University's Peabody College of Education. He is in the Experimental Education Research Training Program (ExpERT) at Vanderbilt, supported by the U.S. Department of Education's Institute for Education Sciences (IES). Mr. Hebert's research interests include writing development, reading development, and how writing may influence reading development, especially for students with reading and writing difficulties. He is the coauthor of the influential meta-analysis *Writing to Read*, funded by Carnegie Corporation of New York. He has several years of classroom teaching experience at the elementary level, including a year teaching on a Navajo reservation in Arizona. Mr. Hebert has several years of experience as a reading specialist in El Segundo, California, where he taught students with reading difficulties, and he is a National Writing Project Fellow through the California Writing Project at UCLA.

Acknowledgments

The authors would like to thank Andrés Henríquez, program officer at Carnegie Corporation of New York, who offered helpful suggestions in conceptualizing and reporting the research reviewed. We also wish to thank Bob Rothman at the Alliance for Excellent Education for his helpful feedback on earlier drafts of this document. In addition, we thank Paul Morphy, who provided technical assistance in calculating some effect sizes, as well as Christopher Lyon, ManFung Lam, and Katie Quille, who helped with various aspects of the project, including conducting electronic searches and locating studies.

Steve Graham
Karen R. Harris
Michael Hebert
Vanderbilt University

CONTENTS

FOREWORD

Timing is everything—or, at least, almost everything. Writing assessment, used formatively to improve learning and instruction, has drawn the attention of researchers and practitioners over the last forty years. Now, at the cusp of the implementation of new assessments measuring the common core state standards in English language arts, Steve Graham, Karen Harris, and Michael Hebert have strategically released their work on formative assessment for writing.

They summarize their extensive ventures into the writing assessment domain with clarity and palpable enthusiasm. What history must their optimism overcome? Over the years, writing assessment research and practice has suffered from dissension at every point, on almost every feature of stimulating, producing, evaluating, and teaching writing. We have argued (empirically and, excuse me, imperially) on goals, procedures, inferences, and links to learning.

One area of multiple fractures derives from a simple question: What should students write about? Information from inside their heads, from assumed common experiences, or from their interpretation of illustrative materials?

And there have been many other divisions: What media should be used? What level of detail might be provided to help students have something to write about? In which of many formats? In imaginative writing, what background knowledge requirements should be met, and which types of domain-relevant prior knowledge are needed in content-based writing? Scholars have argued whether the principal purpose of writing is as a means of expression, communication of knowledge, evidence of comprehension, or audience-focused. I would guess all of the above, although each purpose has different assessment criteria.

Writing researchers and practitioners also have beaten the rubric topic into a moderate level of submission, starting with very general guidance for rating (depending on intuition and "tie-breaking" judgments), moving to a harder but unfeasible set of prompt-specific scoring requirements, and back again to principles or generalizable rubrics, with elements of domain specificity, based in part on research on expert-novice comparisons.

And more: Who scores student writing, and what is their level of preparation? What kind of professional development is needed for self-evaluation, peer rating, teachers' scoring, and scoring by noneducators?

What about the current state of the art in computer scoring? In turn, how can students have sufficient opportunities for writing and feedback in order to develop secure and appropriate skill (and art)? How long should writing segments be? Should there be, within any examination, the opportunity to revise? (The answer is certainly, in a formative assessment setting.) Which of the somewhat formulaic "writing processes" should be promoted, and which discarded?

Finally, the technical quality of the written product is central. Do we care about cognitive processes yielding excellent prose? For the most part, the technical quality space is filled by two topics: 1) the fairness of the prompt for a range of student groups, and 2) the reliability, or level of agreement among raters. Score reliability is rarely treated, and the validity of the writing assessments for a range of intended purposes is left dangling by many researchers. I raise this concern because of the desire to use assessments designed for one purpose to serve another. I can also remember superheated arguments about numbers of raters per paper, percentage of rescoring, audit and moderation processes. (There are more issues, but I have a constraint on words for this foreword.) We are in some way revisiting our goals and lessons learned with undiminished faith that we will produce a nation of writers. Why should we be hopeful, given our track record?

I admire the clear goals, transparent organization, and accessible writing that Graham, Harris, and Hebert have produced. I certainly agree with their general premise about the value of formative assessment in writing. Because no short essay can address every concern, let me comment first on what is in this publication and then on what might be of additional interest.

The authors use "evidence" in their title—but, beyond pronouncements by experts, descriptive studies of students' proficiency or lack thereof, and scarcity in classroom instruction, evidence (by the authors' own admission) is sparse. They bring to bear some research on how writing assessment can be used, citing data on the effects of providing feedback (an integral part of formative assessment), teaching students to assess their own writing, and monitoring progress. In a section on best practices, they suggest letting students choose their preferred medium—paper-and-pencil or computer—in formative assessment. This suggestion only works if accountability measures also permit similar choices. They describe the general issue of construct-irrelevant variance in their recommendations about handwriting, ordering papers for rating, and preference for media. To bolster their findings, they use effect sizes derived from three to sixteen different studies for each topic, with a mode of seven studies providing an effect size.

At issue in all of these studies are the details of the experiments and, most important, the outcome measures used. Standardized assessments are likely to underestimate effects, so the argument for some of the authors' recommendations might be stronger than they suggest. The authors also opine that students should be given multiple opportunities to write, across topics or genre. Without multiple opportunities within genre, though, students are unlikely to master any one. So I would prefer a tight constraint on

genre types and a larger investment in writing across topics. The details of acceptable scorer reliability is open to discussion, and I don't see any clear consensus in this regard, particularly in formative assessment.

And now to what is hard, and awaits another work by these authors. First, teachers need to be taught how to be good formative evaluators—a precondition of which is that they "get" what it is to write, know how to write themselves, understand their students' conceptual and language development, and can communicate productively with them in the time they have available. Given limitations of experience, particularly in elementary and non-"English" classes in secondary school, the professional development problem is daunting, unless clear models, feedback, and adequate time are available.

Second, the sad truth about writing performance derives not exclusively from the National Assessment of Educational Progress findings and persistent differences among identifiable subgroups, but from findings in international comparisons as well. Across a range of topics, the United States has lost more than a step in comparison to many other countries. In writing there is a ray of hope, but the U.S. remains far behind other countries in its ability to close gaps between socioeconomic and immigrant categories across the board.

The authors' next work should include evidence (and precursive research) on how to improve writing for low-performing students without sacrificing the students in the middle or top of the distribution of quality. We also need the features of quality writing to be more carefully articulated through a series of studies on validity. Such an enterprise is essential for students and teachers to support learning, for parents and policymakers, and for researchers to document that assessments are sensitive to additional instructional practices.

If one believes as I do that all writing assessment will soon by delivered and scored on computer or other media, there are additional topics needing evidence. First, researchers must improve the ability to scan optically student handwritten papers to enable computerized scoring of them. Second, automated scoring should be focused on the meaning of students' texts rather than on their structure and word use. Third, the problem of comparability of "long" or scenario-based assessment tasks needs a new psychometric approach. There is always more work to be done, but the authors have—with carefully honed constraints—undertaken clarifying assessment in an important and changing learning domain. May they continue and prosper, along with our students, in the fascinating realm of writing.

Eva L. Baker, EdD
Director, National Center on Research on Evaluation, Standards, and Student Testing
University of California, Los Angeles

EXECUTIVE SUMMARY

The Challenge

Although some progress has been made in improving the writing achievement of students in American schools during the last twenty years (Salahu-Din, Persky, and Miller, 2008), most students do not write well enough to meet grade-level demands. The inability to effectively convey thoughts and ideas through writing plays a role in why many of these students do not complete high school. Among those who do graduate, many will not be ready for college or a career where writing is required. These young people will be at a serious disadvantage in successfully pursuing some form of higher education, securing a job that pays a living wage, or participating in social and civic activities.

The Approach

During this decade there have been numerous efforts to identify instructional practices that improve students' writing. These include *Reading Next* (Biancarosa and Snow, 2004), which provided a set of instructional recommendations for improving writing, and *Writing Next* (Graham and Perin, 2007) and *Writing to Read* (Graham and Hebert, 2010), which were systematic reviews of high-quality research that identified effective writing practices for improving both writing and reading, respectively. Despite these efforts and efforts by others (e.g., Bangert-Drowns, Hurley, and Wilkinson, 2004; Rogers and Graham, 2008), educators and policymakers need additional evidence-based practices for improving the writing of students in American schools.

One tool with potential for improving students' ability to effectively convey thoughts and ideas through text is classroom-based writing assessment. Such formative assessments allow teachers to gauge the effectiveness of their instructional practices, modify instruction as needed, and provide students with feedback on writing strengths and areas in need of further development. These assessments can be administered in a variety of ways in the classroom, including teachers assessing students' writing, students assessing their own writing, and peers assessing others' writing.

This report provides evidence to answer the following two questions:

1. Does formative writing assessment enhance students' writing?

2. How can teachers improve formative writing assessment in the classroom?

This is the first report to examine the effectiveness of formative writing assessment (question 1) using the powerful statistical method of meta-analysis. This technique allows researchers to determine the *consistency* and *strength* of the effects of an instructional practice, and to highlight practices holding the most promise. This report also identifies best practices in writing assessment that need to be implemented in order to maximize the accuracy and trustworthiness of formative writing assessment (question 2).

The Recommendations

1. USE FORMATIVE WRITING ASSESSMENT TO ENHANCE STUDENTS' WRITING

- **Provide feedback.** Writing improves when teachers and peers provide students with feedback about the effectiveness of their writing.

- **Teach students how to assess their own writing.** Writing improves when students are taught to evaluate the effectiveness of their own writing.

- **Monitor students' writing progress.** Writing improves when teachers monitor students' progress on an ongoing basis.

2. APPLY BEST PRACTICES FOR ASSESSING WRITING IN THE CLASSROOM

- **Allow students to use the mode of writing in which they are most proficient when completing a writing assessment.** Writing improves when students are assessed in the format with which they are most experienced—pencil and paper, or word processing.

- **Minimize the extent to which presentation forms such as handwriting legibility or computer printing bias judgments of writing quality.** Writing assessment improves when teachers judge the quality of student writing and do not allow factors such as handwriting or computer printing to bias their judgment.

- **Mask the writer's identify when scoring papers.** Writing assessment improves when teachers do not allow their knowledge of who wrote a paper to influence their judgment.

- **Randomly order students' papers before scoring them.** Writing assessment improves when teachers score papers randomly rather than allow a previous paper's score to influence their judgment.

- **Collect multiple samples of students' writing.** Writing assessment improves when teachers assess students' writing in a variety of genres. This finding supports the decision by the authors of the Common Core State Standards Initiative to emphasize students' mastery of many different types of writing, since writing is not a single generic skill.

- **Ensure that classroom writing assessments are reliably scored.** Writing assessment improves when teachers use procedures for ensuring that particular aspects of writing, such as quality and its basic attributes, are measured reliably.

Informing Writing does not identify all the ways that assessment can enhance writing any more than *Writing Next* (Graham and Perin, 2007) or *Writing to Read* (Graham and Hebert, 2010) identified all possible ways to improve, respectively, students' writing or reading. Nor does it identify all possible best practices in writing assessment. However, all of the recommendations presented in *Informing Writing* are based on strong empirical evidence. The findings are clear: Formative writing assessment makes a difference in how well students convey thoughts and ideas through text. Writing improves when students receive feedback about writing, students evaluate their writing, and teachers monitor students' progress. However, the trustworthiness of formative writing assessments can be compromised if careful attention is not directed at what is assessed, how it is assessed, and how it is scored.

Taken together, the findings from *Writing Next, Writing to Read*, and *Informing Writing* demonstrate that there are a variety of effective instructional methods teachers can apply to improve the writing and reading achievement of students in American schools. The authors hope that in addition to providing classroom teachers with research-supported information about how formative assessment improves writing, this report will stimulate discussion and action at the policy and research levels, leading to the greater use of formative writing assessment in the classroom and the development of better assessment tools.

INTRODUCTION

Skilled Writing Is Essential to Success in the Twenty-first Century

There was a time, not too long ago, when jobs that paid a living wage and required little to no writing on the part of workers were common (Berman, 2009). Today, it is difficult to find such a job. More than 90 percent of white-collar workers and 80 percent of blue-collar workers now indicate that writing is important to job success (National Commission on Writing, 2006).

Between one-half and two-thirds of future jobs in the United States will require a college education (Carnevale and Derochers, 2004; Kirsch, Braun, Yamamoto, and Sum, 2007), and good writing is essential to college success. New college students are expected to be able to write a summary of information from multiple sources, present and defend a point of view in writing, organize information into a coherent written report, and use writing as a tool for learning (ACT, 2005). These are the same writing skills that are needed for success at work (National Commission on Writing, 2004).

The growing demand for better writing skills places new pressures on American schools and students (Nagin, 2003). Youths who cannot effectively convey thoughts and ideas through writing are more likely to receive lower grades, especially in classes where learning is assessed through projects and tests requiring written responses (Graham, 2006). They are also less likely to benefit from the power of using writing as a tool for learning in science, math, social studies, and other classes (Bangert-Drowns et al., 2004; Graham and Hebert, 2010; Graham and Perin, 2007; Klein, 1999). Unless schools help these students write at a higher level, their prospects for finishing school, becoming college and career ready, and obtaining a job that pays a living wage are diminished.

The new grade-level expectations for writing detailed in the Common Core State Standards Initiative (http://www.corestandards.org/the-standards/english-language-arts-standards) provide a road map for the writing skills students need to acquire by the end of high school to be ready for college and a career, emphasizing that writing is not a generic skill but requires mastering the use of writing for multiple purposes. Forty-four states and the District of Columbia have adopted the common core state standards (Gewertz, 2011). These benchmarks are more cohesive and challenging than the writing standards that most states currently apply. Two common-assessment consortia (the Smarter Balanced Assessment Consortium and the Partnership for Assessment of Readiness for College and Careers

(PARCC)), which collectively involve forty-five states and the District of Columbia, are currently developing assessments for measuring these standards at each grade level (Gewertz and Robelen, 2010). These assessments, which are expected to be implemented in School Year 2014–15, will increase the pressure for schools to do a better job of teaching students to write.

The importance of writing extends beyond the classroom and the world of work (DeVoss, Eidman-Aadahl, and Hicks, 2010). Technological innovations have made writing central to social, community, and civic participation in twenty-first-century life. Emailing, blogging, Facebooking, texting, and other electronic writing forms have become ubiquitous means for communicating with family, friends, colleagues, and even people unknown to the writer. In essence, writing has become part of the basic fabric of life in the United States.

Students' Writing Is Not What It Should Be

Good writing is not just an option for young people; it is essential. An inability to effectively use writing to convey thoughts and ideas prevents many American students from completing high school, obtaining a postsecondary degree, acquiring a job that pays a living wage, and participating fully in community and civic life. Although the nation has made some small progress in improving students' writing, too many adolescents are not good writers. According to findings from the latest National Assessment of Educational Progress (NAEP; Salahu-Din et al., 2008), only 33 percent of eighth-grade students and 24 percent of twelfth-grade students performed at or above the "proficient" level (defined as solid academic performance) in writing. In contrast, 55 percent and 58 percent of eighth- and twelfth-grade students, respectively, scored at the "basic" level, denoting only partial mastery of the writing skills needed at their grade level. The rest of the students (12 percent of eighth graders and 18 percent of twelfth graders) scored below the basic level.

Problems acquiring needed writing skills are exacerbated for students who do not speak English as their first language, have a disability, or are black, Hispanic, or Native American. The writing performance of these groups of students on the NAEP was significantly lower than the writing performance of students who were native English speakers, did not have a disability, or were white. The

CAUSE FOR CONCERN

- Poor writing skills cost businesses $3.1 billion annually (National Commission on Writing, 2004).
- Only one out of four twelfth-grade students are proficient writers (Salahu-Din, Persky, and Miller, 2008).
- Nearly one-third of high school graduates are not ready for college-level English composition courses (ACT, 2005).
- College instructors estimate that half of high school graduates are unprepared for college-level writing (Achieve, Inc., 2005).
- College graduates earn 70 percent more than high school graduates (Taggart et al., 2001).
- More than half of adults scoring at the lowest literacy levels are dropouts (National Center for Education Statistics, 2005).

results from the NAEP clearly document that large numbers of adolescents need help to become better writers. As the National Commission on Writing (2003) bluntly states, the writing of students in the United States "is not what it should be (p. 7), and "[w]e must do better" (p. 17).

Solutions

School success and college and career readiness require that all young people possess strong writing skills. They must be able to write in an accurate and clear manner and use writing to inform and persuade others (ACT, 2005; Berman, 2009; National Commission on Writing, 2004, 2005). During the past decade, considerable effort has been made to improve literacy instruction for younger children as well as for youths in middle and high school. Much of this work has focused on reading (e.g., Biancarosa and Snow, 2004; National Institutes of Children's Health and Development, 2000; Scammacca et al., 2007), with much less attention given to writing. As the National Commission on Writing (2003) notes, writing "is clearly the most neglected" of the three Rs (p. 3). It is past time to give greater attention to another equally important chapter in the literacy reform movement.

The groundwork for improving writing instruction in American schools has been constructed by many organizations and individuals. Several reviews of high-quality research summarize various aspects of effective writing instruction. For instance, *Reading Next* (Biancarosa and Snow, 2004) identified writing as a key element in good literacy instruction. *Writing Next* (Graham and Perin, 2007) demonstrated that there are many effective tools for teaching writing (see also Graham and Perin, 2007; Rogers and Graham, 2008). This same report and a previous review by Bangert-Drowns et al. (2004) also verified that writing about ideas in science, social studies, and other content classes enhances learning. *Writing to Read* (Graham and Hebert, 2010) confirmed that students comprehend text better if they write about it, and that students become better readers as a result of receiving writing instruction. Other organizations have been focused on improving the implementation of effective writing instruction and practices. For the past forty years, the National Writing Project (http://www.nwp.org/) has helped American teachers become better teachers of writing. The National Commission on Writing (2003) provided a blueprint for reforming writing instruction, recommending that writing and writing instruction time be doubled, teacher preparation be improved, and technologies for teaching, producing, and assessing writing be applied. This blueprint also emphasized the critical role of assessment in reforming writing instruction, noting that "individual students need to know their strengths and weaknesses, [and] their teachers also need to understand when students are writing effectively and when they are experiencing difficulty" (p. 21).

Assessing Writing

This report focuses on the assessment of writing. Writing assessment occurs for many different purposes. Teachers assess writing to monitor students' progress, inform instruction, provide feedback, and judge the effectiveness of their teaching. Students assess their own writing to appraise growth,

determine strengths, and identify areas in need of further development. Peers assess each other's writing, providing feedback on what works and what still needs improvement. Schools assess writing to determine how many students meet local or state performance standards and identify youths who need extra help. States and the federal government administer writing tests to measure American students' collective writing success, evaluating students' ability to effectively convey thoughts and ideas through writing across time. Employers assess writing to make decisions about whom to hire and promote.

Currently, the most visible and influential assessments of students' writing involve efforts to determine how many children meet local or state performance standards. Most states conduct such summative, high-stakes assessments once a year in selected grades (Beck and Jeffery, 2007; Jeffery, 2009). Unfortunately, states and school districts implemented these tests before adequate evaluations of the consequences of such assessments were conducted. There is now limited evidence (mostly qualitative and correlational) that these assessments make writing instruction more central to the mission of schools, change teachers' writing practices in positive ways, and improve students' writing (Callahan, 1999; Dappen, Isernhagen, and Anderson, 2008; Parke, Lane, and Stone, 2006). The value of such tests has been questioned, as there is evidence that the measures narrow the teaching of writing to only what is assessed (Hillocks, 2002), limit originality and voice as students are taught formulaic approaches for writing in the genres tested (Albertson, 2007), and send the message that writing is solely the job of language arts teachers, since writing in specific content areas is rarely assessed (Callahan, 1999). Concerns have further been raised about the fairness of such assessments, as analyses of state writing tests reveal that what is assessed, how it is assessed, and how it is scored vary from state to state (Jeffery, 2009), and such differences have direct consequences for how many students receive a passing or failing score (Haladyna and Hess, 1999–2000; Moon and Hughes, 2002; Popp and Ryan, 2003). Because the scoring of these tests is time-consuming, teachers and students often must wait months before results are available, limiting the value of the assessment information and feedback.

The two common-assessment consortia, which are now designing assessments for the new set of common core academic standards (Gewertz, 2011), will likely solve some of the problems that plague summative, high-stakes writing tests—such as, for example, the need for greater uniformity—but not all of them. A potentially important aspect of the work of these consortia is the development of formative writing assessments. These involve assessments that provide up-to-date information or feedback about students' progress, allowing teachers, students, or both to adjust what they are doing. For instance, teachers can monitor students' ongoing progress in writing along several dimensions, such as ideation, organization, word choice, and mechanics (spelling, grammar, and usage). This allows teachers to gauge the effectiveness of their instructional practices, modify instruction as needed, and provide students with feedback on writing strengths and areas in need of further development in each of these parts.

While teachers report that they use formative measures to evaluate students' writing, relying on scoring rubrics, professional judgment, and other forms of formative assessment, including peer and self-

evaluation, most teachers indicate that they use these assessment procedures infrequently (Gilbert and Graham, 2010; Graham, Capizzi, Hebert, and Morphy, 2010; Kiuhara, Graham, and Hawken, 2009). The development of new formative writing assessments by the two common-assessment consortia is likely to increase teachers' use of these types of measures. Consequently, it is important to determine whether the use of such assessments in the classroom makes a meaningful difference in how well students write.

This report provides evidence to answer the following questions:

1. Does formative writing assessment enhance students' writing?
2. How can teachers improve formative writing assessment in the classroom?

Formative writing assessment can take many forms. It can involve assessments conducted by the teacher, the student, or classroom peers. This report examines whether all of these forms of assessment are effective. Formative writing assessments can be biased and even invalidated by issues involving what is assessed, how it is assessed, and how it is scored. This report also identifies issues that teachers need to address to ensure that formative writing assessments provide an accurate and trustworthy measure of how effectively students convey thoughts and ideas through writing.

RECOMMENDATIONS FOR USING FORMATIVE WRITING ASSESSMENT TO IMPROVE WRITING

Assessment is commonly recommended as a means for improving writing instruction. Assessments must be conducted with great care, however, as all tests have consequences. The types of assessments that teachers typically undertake influence what and how writing is taught, what kind of feedback students receive about their writing, and which students get extra help from the teacher. Because assessment is evaluative, teacher assessments impact students' grades and perceptions of their writing competence (Andrade, Wang, Du, and Akawi, 2009). As a result, assessment should not be entered into lightly and must be based on best practices.

This report provides long-needed guidance for teachers, schools, and policymakers on the impact of formative writing assessment and the factors that influence such assessments. The special contribution of the report is that the recommendations are based on scientific evidence collected in grades one to twelve. The findings and recommendations drawn from this literature apply to the classroom as well as the formative writing assessment procedures being developed for the Common Core State Standards Initiative (CCSSI) (Gewertz, 2011).

To answer the first question—"Does formative writing assessment enhance students' writing?"—the authors set out to collect, categorize, and analyze data from true-experimental and quasi-experimental studies (see Appendix A for details). In both types of studies, students in an experimental group receive a specific intervention (formative writing assessment, for instance), and their performance is compared to a control group of students that receives a different treatment or no treatment. True-experimental studies control for preexisting differences between students in the two groups through random assignment, whereas quasi-experimental studies do so through other means, such as administering a pretest so that preexisting differences can be controlled. For each study, an effect size was calculated. An effect size provides a measure of the direction and magnitude of the difference between the treatment and control condition. Meta-analyses were then conducted, which provided an average weighted effect size across studies for each of the different types of formative writing assessments examined (Lipsey and Wilson, 2001). These are the first meta-analyses examining the effects of formative assessments on students' writing performance.

APPENDIXES

Appendix A describes the methodology used to locate and categorize studies.

Appendix B lists the studies and presents the data used in this report.

A TECHNICAL NOTE ON META-ANALYSIS

What Is Meta-analysis?

Meta-analysis is a particularly powerful way of summarizing empirical research, as it aggregates the findings from studies by calculating an effect size for each one. The strength of meta-analysis is that it allows consideration of both the strength and consistency of a treatment's effects.

What Is an Effect Size?

An effect size reports the average difference between one type of instruction (or condition) and an alternative or control condition. It indicates the **strength** of the difference between the two conditions. The following guidelines provide a benchmark for interpreting the magnitude of an effect:

0.20 = **small** or mild effect

0.50 = **medium** or moderate effect

0.80 = **large** or strong effect

A **positive** effect size means that the experimental treatment or condition had a positive effect on students' writing when compared to the control condition.

A **negative** effect size means that the control condition had a stronger effect on students' writing than the experimental treatment or condition.

To answer the second question—"How can teachers improve formative writing assessment in the classroom?"—a broader set of empirical data was collected, categorized, and analyzed. This data included true- and quasi-experimental studies, investigations where students were their own controls (receiving both the treatment and control condition), correlational studies where the relationship between students' performance on different writing tasks was examined, and investigations examining the reliability of different formative writing assessment measures.

Recommendations

1. USE FORMATIVE WRITING ASSESSMENT TO ENHANCE STUDENTS' WRITING

- **Provide feedback.** Writing improves when teachers and peers provide students with feedback about the effectiveness of their writing.
- **Teach students how to assess their own writing.** Writing improves when students are taught to evaluate the effectiveness of their own writing.
- **Monitor students' writing progress.** Writing improves when teachers monitor students' progress on an ongoing basis.

2. APPLY BEST PRACTICES FOR ASSESSING WRITING IN THE CLASSROOM

- **Allow students to use the mode of writing in which they are most proficient when completing a writing assessment.** Writing improves when students are assessed in the format with which they are most experienced—pencil and paper, or word processing.

- **Minimize the extent to which presentation forms such as handwriting legibility or computer printing bias judgments of writing quality.** Writing assessment improves when teachers judge the quality of student writing and do not allow factors such as handwriting or computer printing to bias their judgment.

- **Mask the writer's identity when scoring papers.** Writing assessment improves when teachers do not allow their knowledge of who wrote a paper to influence their judgment.

- **Randomly order students' papers before scoring them.** Writing assessment improves when teachers score papers randomly rather than allow a previous paper's score to influence their judgment.

- **Collect multiple samples of students' writing.** Writing assessment improves when teachers assess students' writing in a variety of genres. This finding supports the decision by the authors of the CCSSI to emphasize students' mastery of many different types of writing, since writing is not a single generic skill.

- **Ensure that classroom writing assessments are reliably scored.** Writing assessment improves when teachers use procedures for ensuring that particular aspects of writing, such as quality and its basic attributes, are measured reliably.

1. USE FORMATIVE WRITING ASSESSMENT TO ENHANCE STUDENTS' WRITING

The evidence demonstrates that including formative writing assessment as part of classroom instruction enhances students' writing. The positive effects of formative assessment occur when teachers or peers provide students with feedback about their writing or the learning of a particular writing skill, students are taught to assess their own writing, and teachers monitor students' progress on an ongoing basis.

Provide Feedback

Average Weighted Effect Size = 0.77
Based on sixteen comparisons where the outcome was students' overall quality of writing

A long-term staple of writing instruction is for teachers to provide students with feedback about one or more aspects of their writing. Feedback is not just limited to teacher comments about a paper, though. It can involve comments about students' progress in learning writing skills or strategies,

EXAMPLES OF FEEDBACK

- Teacher gives students specific feedback on the progress they are making in learning a strategy for writing a paragraph (Schunk and Swartz, 1993a).
- Students are taught a strategy for giving feedback to their peers on substantive (e.g., clarity and completeness) and mechanical issues (e.g., misspelling and usage). Students provide and receive feedback to and from each other using the strategy (MacArthur, Schwartz, and Graham, 1991).
- The writer observes another student carry out directions he or she wrote, and then revises the directions based on the observation (Couzijn and Rijlaarsdam, 2005).
- Students receive feedback weekly from the teacher on their writing productivity—the amount written and spelling accuracy (Rosenthall, 2006).

responses from a parent about a written composition (written or verbal comments as well as watching someone enact what was written), reactions from peers (with or without instruction on how to do so), or some combination of these options. When considered together, these forms of feedback had a positive and statistically significant impact on how well students conveyed thoughts and ideas through writing (the effect size confidence interval ranged from 0.49 to 1.08). In all sixteen studies, feedback had a positive effect. It is important to note, however, that these findings extend only from grades two to nine (see Appendix B, Table 1).

The two forms of feedback that were most studied by researchers in the sixteen studies were examined. First, when peers both gave and received feedback on their papers to and from each other, a statistically significant average weighted effect size of 0.71 was obtained across six studies (the effect size confidence interval ranged from 0.29 to 1.14).

Second, when peers just received feedback from an adult or a peer, a statistically significant average weighted effect size of 1.01 was obtained across eight studies (the effect size confidence interval ranged from 0.48 to 1.55). The types of feedback tested in these studies mostly involved teacher feedback about students' progress in learning writing skills or processes (three studies) and verbal feedback from an adult or peer about their text (three studies). Consequently, a claim cannot be made about the effectiveness of the most common conception of feedback in writing—teachers providing students with written comments about one or more aspects of their writing. Only one of the experimental and quasi-experimental studies included in our analysis assessed this form of feedback, yielding an effect size of 0.29 (Rosenthall, 2006).

Teach Students How to Assess Their Own Writing

Average Weighted Effect Size = 0.46
Based on seven comparisons where the outcome was students' overall quality of writing

EXAMPLE OF TEACHING STUDENTS TO EVALUATE THEIR OWN WRITING

Students read and discuss a model essay, discuss its strengths and weaknesses, and develop a list of the qualities of a good essay. The teacher presents a rubric for scoring essays and describes and shows how to use it. The rubric assesses seven attributes of students' writing: ideas and content, organization, voice, word choice, sentence fluency, and conventions. The score for each attribute ranges from 0 to 3, and a description of what the papers should do is provided for each score. Students use the rubric to score the first draft of a paper prior to revising it (Andrade, Wang, Du, and Akawi, 2009).

Teaching students how to assess their own writing has a positive and statistically significant effect on how effectively students convey thoughts and ideas through writing (the effect size confidence interval ranged from 0.07 to 0.94). Self-evaluation procedures ranged from teaching students to use a rubric to assess the merits of specific features of their writing (e.g., ideation, organization, voice, vocabulary, sentence formation, and conventions) to teaching specific strategies for evaluating a first draft of a paper for substantive (e.g., clarity) or mechanical (e.g.,

misspelled words) lapses to teaching students how to detect mismatches between what they intended to say and what they wrote.

Self-evaluation of writing had a consistently positive impact on the quality of students' writing, as six of the seven comparisons (86 percent) produced a positive outcome (see Appendix B, Table 2). These gains occurred for students in grades three to twelve.

Monitor Students' Writing Progress

Average Weighted Effect Size = 0.24
Based on published standardized norm-referenced tests and state writing tests from seven comparisons

When teachers assess or monitor students' writing progress, it has a positive and statistically significant impact on students' overall writing performance (the effect size confidence interval range was from 0.03 to 0.45). All of the outcomes for the seven available studies were positive (see Appendix B, Table 3) and involved treatments ranging from teaching teachers how to assess students' papers in terms of ideas, organization, voice, and usage/conventions to frequently collecting information on how much students write as well as on the overall correctness of their compositions for word choice, grammar, usage, and spelling. The findings for monitoring students' writing progress mostly involve weaker writers and students with special needs, and are limited to the elementary and middle school grades. Thus, the effectiveness of monitoring writing with older students is unknown, and there is only meager evidence on its impact with typically developing writers.

The form of monitoring students' writing progress that was most prominent in the seven studies reviewed was curriculum-based assessment. With this approach, teachers assess students' writing frequently to determine if the class and individual students are making adequate progress, and to adjust their writing instruction accordingly (Espin, Weissenburger, and Benson, 2004). Typically, a sample of students' writing performance is collected weekly, scored, and graphed. In each assessment, students write for the same amount of time. Because teachers must score what each student writes, time for writing is generally limited to no more than five minutes (although this is not essential).

2. APPLY BEST PRACTICES FOR ASSESSING WRITING IN THE CLASSROOM

The finding that formative writing assessment enhanced students' writing must be tempered by the challenges of assessing writing in the classroom. If such assessments are to be valid and fair, they must be based on best practices. Six best practices derived from meta-analyses, investigations examining the relationship between students' performance on different writing tasks, and reliability studies are presented next.

CURRICULUM-BASED ASSESSMENT MEASURES

"Correct Word Sequences": the number of combinations of two adjacent, correctly spelled words that make sense and are the correct part of speech given the context of the sentence (correct punctuation or capitalization is counted as a correct word). Variations of this measure are to calculate the percentage of correct word sequences; subtract the number of incorrect words from the number of correct word combinations; and determine the average number of correct combinations that occur before there is an incorrect word.

"Words": the total number of words written (a word is counted even if it is misspelled or grammatically incorrect).

"Words Spelled Correctly": the number of legible words that are spelled correctly. A variant of this score is percentage of correctly spelled words.

Allow Students to Use the Mode of Writing in Which They Are Most Proficient When Completing a Writing Assessment

Allow Students Experienced in the Use of Word Processing to Write Electronically When Tested

Average Weighted Effect Size = 0.54
Based on seven comparisons where the outcome was students' overall quality of writing

In *Writing Next*, Graham and Perin (2007) reported that students who used word processing as their primary mode of composing were better at conveying thoughts and ideas through text than students who wrote by hand (the average weighted effect size was 0.50). In this report, a similar effect was found during the assessment of writing. In seven true and quasi-experiments, where one group of students wrote using word processing during testing and the other group wrote by hand (see Appendix B, Table 4), a statistically significant effect of 0.54 favoring word processing was obtained (the effect size confidence intervals for our analysis ranged from 0.49 to 1.08). Thus, composing by hand when taking a writing test may underestimate a student's ability to effectively convey thoughts and ideas through writing, providing a biased estimate of the student's writing capabilities.

Allow Students to Write Using Paper and Pencil When They Are Inexperienced in the Use of Word Processing

Average Weighted Effect Size = 0.48
Based on three comparisons where the outcome was students' overall quality of writing

The situation is more nuanced than the findings from the seven studies above would suggest, however. Whereas all seven comparisons produced a positive effect favoring word processing, one study produced a particularly small effect (0.04). This investigation by Russell and Plati (2000, study 1) mostly involved students with little word processing experience. Three additional studies were located (see Appendix B, Table 4) where students with more, little, or no word processing experience wrote compositions by hand and on the word processor during testing. These studies did not involve a control group; rather, students served as their own controls. That is, when students with little to no word processing experience in these three studies wrote by hand, they produced compositions that were judged to be statistically better than the ones they produced when writing on a word processor

(the average weighted effect size was 0.48). For these students, taking a writing test using a word processor underestimated their writing capabilities. Thus, students must be matched to the appropriate mode of composing if formative writing assessments are to provide an accurate evaluation of students' capabilities to effectively convey thoughts and ideas through writing.

Minimize the Extent to Which Presentation Forms Such as Handwriting Legibility or Computer Printing Bias Judgments of Writing Quality

A long-standing concern when assessing writing is that judgments about the quality of a writer's message are influenced by factors involving the physical presentation of the composition (James, 1927). Although most people agree that good writing is legible, the analyses demonstrated that poor legibility can exert an excessive negative influence on assessment of students' ability to effectively convey thoughts and ideas through writing.

Less Legible Versus More Legible Handwritten Text

Average Weighted Effect Size = -1.03
Based on five comparisons where the outcome measure was writing quality

When assessing students' writing, teachers and teachers in training score the thoughts and ideas in a less legible version of a paper more harshly than they score a more legible version of the same paper. (The effect size confidence interval ranged from -1.40 to -0.65.) This statistically negative effect was based on studies with students in grades six to twelve, and a negative effect for less legible handwriting was observed in all five comparisons (see Appendix B, Table 5). While teachers should clearly be aware of the effects of legibility, it is not known if such knowledge minimizes its effects. One way of eliminating presentation effects due to legibility is to type all papers before scoring them (Graham, 1990). Because of the time that this requires, however, teachers are unlikely to adopt this solution, except on a very limited basis. It must be noted that legibility is not an issue if assessments are completed on a word processor, but as the next analyses demonstrated, scoring typed text is not without problems.

Computer Versus Handwritten Text

Average Weighted Effect Size = -0.47
Based on five comparisons where the outcome measure was writing quality

Teachers give statistically lower overall scores for writing quality to a printed computer copy of a composition than they do to an exact handwritten copy of the same paper (the effect size confidence interval ranged from -0.79 to -0.27). The negative effect for the typed test was consistent across all five studies, which were conducted with students in grades four to ten (see Appendix B, Table 5). According to Russell and Tao (2004b), scorers are harsher in scoring the thoughts and ideas included on a computer copy of text because errors are more visible than they are on a handwritten copy. One way of minimizing the biasing effect of computer text on scoring students' written thoughts and ideas

is to correct spelling, grammar, and usage errors before scoring. Given the time required to make such corrections, however, teachers are unlikely to employ this solution except in rare circumstances. Other possible solutions for addressing this problem would be to provide teachers with training on the biasing effect of this form of presentation or using a font for computer text that resembles handwritten text. Russell and Tao (2004b) found that both of these methods reduced the negative biasing effect of scoring computer-printed text.

Mask the Writer's Identity When Scoring Papers

It is commonly assumed that knowing something about who wrote a particular paper can influence a teacher's judgment about the quality of the text. For example, a teacher may be inclined to give a higher score for overall text quality than is deserved to a student who has written especially good essays in the past (or a lower score to a student who has crafted a poorly written paper in the past). Likewise, some teachers may make a priori assumptions about the writing of students based on their gender, disability status, or ethnicity, and score these students' writing accordingly. Analysis of available evidence supports this assumption.

Knowledge of Gender, Disability Status, and Ethnicity

Average Weighted Effect Size = -0.58
Based on five comparisons where the outcome was students' overall quality of writing

The findings from the five studies that were reviewed (see Appendix B, Table 6) indicate that knowing something about the writer's identity statistically biased the judgments of teachers and teachers in training on writing quality (the effect size confidence interval ranged from -0.96 to -0.21). An obvious solution to this problem is for teachers to assess each student's writing samples without looking at who wrote it. A better, but more time-consuming, solution is to remove any identifying information about the writer from the paper. Graham and Leone (1987) also found that the biasing effect of knowing something about the writer was minimized by simply providing teachers with training in how to score essays (Graham and Leone, 1987).

Randomly Order Students' Papers Before Scoring Them

An old maxim is that you are known by the company you keep. Based on the findings from a small number of studies, this appears to be the case for writing, too. Two studies (Hales and Tokar, 1975; Hughes, Keeling, and Tuck, 1980) report that an average essay gets a lower-quality score if it is preceded by a set of good essays versus a set of weaker ones (effect size of -0.72 in Hughes et al., 1980, and -0.31 in Hales and Tokar, 1975; see Appendix B, Table 7). Teachers can minimize such context effects by randomly ordering papers before scoring them.

Collect Multiple Samples of Students' Writing

We infer how well students write by examining their performance on one or more writing tasks. Because writing serves many purposes and different forms of writing require specialized knowledge (e.g., narrative versus persuasive), it is unlikely that a single piece of writing provides enough evidence to make reliable and valid judgments about a student's writing across different genres or tasks. This viewpoint is reflected in the new CCSSI, where writing is not viewed as a single generic skill and students are expected to learn to use different types of writing for multiple purposes.

If a single paper provides an adequate measure of writing, then students' scores on different types of compositions (e.g., persuasive, descriptive, expository, narrative) should be similar when they are scored in the same way. The evidence from this report did not support this basic assumption. The findings from the seven studies reviewed reveal that the quality of writing for students in grades two to twelve differed across genres. In each study, the overall quality of students' writing in two or more genres of writing differed statistically (see Appendix B, Table 8). Thus, students' performance on one type of writing is not identical to their performance on another.

Similarly, if a single writing task is a sufficient indicator of students' writing, performance on one writing sample should be highly correlated with performance on other writing samples. Again, this assumption was not supported. In nine studies with students in grades two to twelve (see Appendix B, Table 9), correlations between writing quality scores for different types of compositions (e.g., persuasive, descriptive, expository, narrative) were small to moderate ($r = 0.10$ to 0.60). Even when students wrote multiple papers in a single genre (two studies), the correlations were still primarily moderate in magnitude, ranging from 0.69 to 0.79 in two studies (see Appendix B, Table 9). Consequently, a single piece of writing is not a valid measure of students' performance on the same or different writing tasks. Importantly, these findings support the decision by the authors of the CCSSI to emphasize multiple genres and writing purposes versus treating writing as a single generic skill.

These findings demonstrate that teachers cannot make sound decisions about students' writing or their progress as writers when such decisions are based on a single piece of writing. Likewise, classroom assessments cannot be limited to a single type of writing. One way to address these issues is to collect multiple writing samples when carrying out an instructional unit in a specific genre (such as persuasive writing). Students can keep a portfolio of their persuasive writing during the unit, which can be evaluated by the teacher, students, or the students' peers. Another possible solution is the application of a curriculum-based approach, where students' writing performance in a particular domain is assessed and monitored weekly.

Ensure That Classroom Writing Assessments Are Reliably Scored

A basic assumption underlying any valid assessment is that it is reliable—that is, that students' scores will not change appreciably if their papers are rescored. This assumption is generally valid for writing measures that can be objectively defined and easily counted. Nine curriculum-based writing measures (e.g., total words written, number of sentences, correctly spelled words, correct word sequences) can be reliably scored (see Appendix B, Table 10, for a definition of each measure, and the reliability of each measure in Table 11). A measure was considered reliable if it met the following criteria: exact agreement between scores of 70 percent (consensus approach), and correlation between scores of .80 or greater (consistency approach). In all thirty-six studies, with each measure, the criteria for consensus and consistency were met.

MEASURING RELIABILITY

The **Consensus Approach** calculates the **percentage of exact agreement** to indicate how often raters assign the exact same score.

The **Consistency Approach** calculates a **reliability coefficient** (or correlation) to provide an estimate of the degree to which the pattern of high and low scores among raters is similar.

An exact percentage of agreement of 70 percent or better indicates reliable scoring with the consensus approach (Brown, Glasswell, and Harland, 2004). A reliability coefficient of .80 is generally viewed as acceptable with the consistency approach (Nunnally, 1967; Shavelson and Webb, 1991), but higher coefficients are desirable when scores are used to make decisions about individual students (e.g., a reliability coefficient between .90 and .95).

For measures that are more subjective (i.e., harder to define objectively and not easily counted), scores are not always reliable. The two most common ways of measuring writing quality are holistic and analytic. With a holistic scale, a single rating of general quality of the composition is made, whereas an analytic scale produces separate ratings for specific attributes such as ideation, organization, style, and so forth. Twenty-two studies were located that examined the reliability of holistic scales on everything from high-stakes writing assessments to more typical classroom assessments, including portfolios. In each of these studies there were more than two raters, increasing the generalizability of the findings. Only 25 percent of studies met the 70 percent criteria for consensus, whereas 47 percent of the studies met the .80 criteria for consistency (see Appendix B, Table 12). Twenty-one similar studies that examined the reliability of analytic scales were found. None of the studies met the 70 percent criteria for consensus; only 26 percent of the studies met the .80 criteria for consistency (see Appendix B, Table 13).

The findings from the studies reviewed in this report make it clear that care must be given to establish the reliability of more subjective writing measures, such as holistic and analytic writing scales, if teachers are to use these measures in the classroom. Otherwise, scores from these measures will be too elastic for teachers to make sound decisions about students' writing or their progress as writers.

(Evidence-based practices for improving the reliability of these measures are presented in the sidebar below, and the studies supporting each recommendation are referenced in Appendix B, Table 14.)

IMPROVING RELIABILITY

Reliability can be improved by

- providing training on how to score compositions;
- having multiple teachers score each paper to establish reliability as well as having them discuss and resolve differences in their scores;
- basing students' writing score on multiple writing tasks;
- increasing the scoring range (e.g., instead of a scale with 6 points, using one with 20 points);
- providing teachers with benchmarks (descriptions or examples) for each point on the scale; and
- applying a two-step scoring process where the teacher matches the composition to the closest benchmark, and then scores it again if it does not match this benchmark perfectly by adding a plus or minus to the first score.

Note: Studies supporting these recommendations are presented in Appendix B, Table 14.

IMPLEMENTING THE RECOMMENDATIONS

Much of the groundwork needed for improving writing instruction in American schools has been established. *Reading Next* (Biancarosa and Snow, 2004) identified writing as a key element in good literacy instruction. *Writing Next* (Graham and Perin, 2007) and *Writing to Read* (Graham and Hebert, 2010) verified that writing about material read or presented in science, social studies, and other content classes enhances comprehension and learning. These two reports also demonstrated that there are many effective tools for teaching writing, and that teaching writing enhances students' reading abilities. The National Writing Project (http://www.nwp.org/) has developed an extensive nationwide infrastructure for helping teachers become better teachers of writing. The National Commission on Writing (2003) provided a blueprint for reforming writing instruction, emphasizing, among other things, the critical role that assessment plays in good writing instruction.

This report extends and strengthens the bedrock laid by these previous efforts, providing empirical support that formative writing assessment in the classroom makes a difference. When teachers monitor students' progress, writing improves. When students evaluate their own writing, writing improves. When students receive feedback about their writing, writing improves. When students are partners in writing assessment, giving and receiving peer feedback, students' writing improves. These findings provide empirical support for the decision to develop and include formative writing assessments as a central part of current efforts to assess student progress toward meeting the newly developed Common Core State Standards Initiative (CCSSI) (Gewertz, 2011). The findings also provide empirical support to teachers' past, current, and future use of such assessments in the classroom.

Formative writing assessments have many advantages. They allow teachers to gauge the effectiveness of their instructional practices, modify instruction as needed, and provide students with feedback on writing strengths and areas in need of further development. The value of such assessments is diminished, however, if they are biased or invalidated by issues involving what is assessed, how it is assessed, and how it is scored. This report identifies empirically supported best practices in writing assessment that need to be addressed if such assessments are to provide accurate and trustworthy measures of students' ability to effectively convey thoughts and ideas through writing.

First, formative writing assessments need to be based on multiple writing samples, as students' performance within or across genres cannot be accurately reflected through a single piece of writing. This finding supports the emphasis in the CCSSI that writing is not a generic skill but involves the mastery of using different types of writing for different purposes. Second, students who are experienced and skilled in using word processing to write should be given the option of using this mode of composing when completing formative writing assessments, as their ability to effectively convey thoughts and ideas is likely to be weaker when writing by hand. Conversely, handwriting would be the preferred mode of composing for students with little experience or skills in using word

processing. Third, teachers need to control factors that are likely to bias their judgments when formative assessments involve scoring writing quality or its attributes (e.g., organization, voice, sentence fluency). This includes minimizing the influence of the physical attributes of text (e.g., handwriting legibility or computer-printed text), masking the identity of the writer, and randomly ordering students' writing samples before scoring them. Fourth, care must be taken to ensure that formative assessments are reliably scored by teachers. This is primarily an issue when formative writing assessment involves scoring writing quality or its attributes.

One challenge in implementing formative writing assessments in the classroom is that scoring is labor intensive. As the findings from this report demonstrate, students can help share this load. Writing improves when students evaluate their own writing and when peers give feedback to each other. Such student evaluations are likely to be most effective when students are taught how to carry out these assessments. Moreover, teachers often think that they need to give extensive written feedback to students on their writing. This practice is questionable, as students may be discouraged by extensive written comments and suggestions (Graham, 1982), and evidence was not available to support this practice. This is not to say that teachers should forgo giving written feedback on what students write. Rather, this feedback should be more selective and less extensive.

Another potential means for reducing the scoring load of formative writing assessments is to apply computer scoring systems. Today, machines provide feedback more frequently to writers than other humans do. Whenever something is written on a word processor, one or more automated scoring systems are activated. The most obvious feedback involves spelling, as words that are misspelled (and not corrected by the machine) are highlighted in one way or another. Most word processing programs also contain other automated scoring systems that provide feedback on grammar as well as the number of characters, words, paragraphs, and pages written.

During the last five decades, other more complex computer marking systems have been developed. These systems range from ones where students receive feedback on multiple linguistic features and errors in their text (e.g., *Critique*, from the Educational Testing Service) to computer programs that compare the semantic similarities between one piece of writing and another, such as a student's summary of text read (latent semantic analysis). As part of this report, four experimental studies where students received feedback about their writing from a computer marking system were located (see Appendix B, Table 15). While the overall weighted effect size for these studies was not statistically significant, the combined effect was 0.29 for writing quality/content (confidence intervals ranged from -0.04 to 0.62). Effect sizes above 0.25 are often interpreted as showing practical significance. Consequently, teachers should be encouraged to explore the use of computer marking systems with their students. These procedures ease the scoring burden, are just as reliable at assessing quality as human scorers (Coniam, 2009), and will likely improve with time.

A second challenge in implementing formative writing assessments more fully into the classroom involves teacher preparation. Many teachers, but especially content-area teachers, indicate that their preparation to teach writing is inadequate (e.g., Kiuhara et al., 2009). If such writing assessment practices are to become a positive and routine aspect of classroom life, then the professional development that teachers receive at college, through their schools or school district, and from their profession must provide them with the knowledge and skills needed to implement them effectively.

A third challenge for teachers and schools involves implementing formative writing assessments so that they are applied conjointly with best practices in writing assessment. Teachers, schools, and the common core assessment consortia (Gewertz and Robelen, 2010) should be urged to consider how to put these best practices into action. To illustrate, the actualization of several best practices could involve asking students to put their name on the back of their formative writing samples (reducing the chance that the teacher immediately knows the author), randomly ordering papers before grading them (to reduce context scoring effects), and using rubrics to score papers (to improve scoring consistency).

Finally, formative assessment and best practices in writing assessment hold their greatest promise for helping teachers and schools create students who are skilled and confident writers if they are implemented as part of a comprehensive reform of writing instruction. To address the long-standing concerns about students' writing and the neglect of writing instruction in many classrooms, schools, school districts, and states need to develop new and better policies that establish unambiguous, challenging, and realistic plans for improving writing instruction and students' writing. These plans must establish clearly specified methods and incentives to ensure that

- teachers are prepared to teach writing effectively;
- the teaching of writing is the responsibility of all teachers;
- students write frequently;
- students use writing as a tool to support learning across the curriculum;
- sufficient time is devoted to the teaching of writing at all grade levels;
- schools move from students writing mostly by hand to composing with a variety of tools, including paper and pencil, word processing, and other new technologies for composing; and
- formative and summative assessments provide reliable, valid, fair, and useful information to students, teachers, parents, and policymakers.

A RESEARCH AGENDA FOR THE FUTURE

During the last decade, a number of systematic reviews of the writing intervention literature for students in grades one to twelve have been conducted (e.g., Andrews et al., 2006; Bangert-Drowns et al., 2004; Graham and Hebert, 2010; Graham and Perin, 2007; Rogers and Graham, 2008). These reviews provide strong support for the importance of teaching writing and guidance on how to do so, but each was based on fewer than 125 studies. While the recommendations in the current report draw on 136 studies, research in writing pales in comparison to other academic areas such as reading (Graham and Perin, 2007; National Commission on Writing, 2003). The coming years must bring increased efforts to conduct new research on writing development, instructional practices, and assessment procedures. Such research is dependent on funding; if the field of writing research is to move forward in any significant way, federal and private agencies must make writing research a priority.

Regarding writing assessment, there are a number of gaps in the research base and areas where more evidence is needed. The field needs to develop a research agenda that will strengthen the knowledge base about writing assessment. In particular, there is a need for the development of new formative and summative assessments that are reliable, valid, and fair, as well as methods for determining how such assessments can best enhance writing instruction and students' writing development. A good start would be for federal agencies to fund programs of research for writing assessment similar to the "Reading for Understanding" grants solicited in SY 2009–10 by the Institute for Educational Sciences. This competition was designed to increase the understanding of how to effectively assess reading comprehension. Similar research is needed in writing.

It is hoped that this report will spur new research efforts on writing assessment, especially in the areas listed below.

- *Understanding better how writing develops.* The development of good writing assessments requires a strong understanding of writing development. The current understanding is fragmented and incomplete.

- *Designing and testing new approaches for measuring writing.* For instance, Rijlaarsdam and his colleagues (in press) reported that up to 80 percent of the variability in the quality of students' writing can be accounted for by the timing and types of cognitive activities (e.g., goal setting, generating ideas, evaluating text, rereading text) they engage in while composing. This suggests that a promising avenue for developing new writing assessments involves assessments that move beyond the writing product to consider writing processes as well.

- *Continuing research on computer systems for scoring writing.* Formative and summative writing assessments are time-consuming and expensive. Computer marking systems provide a partial solution to these constraints. They may also help to provide insight into human scoring (see Ben-Simon and Bennett, 2007), possibly leading to improvements in scoring validity or reliability.

- *Identifying appropriate testing accommodations for students with disabilities and English language learners.* There has been surprisingly little research focused on testing accommodations in writing.

- *Studying factors that promote and hinder teachers' use of writing assessment.* Very little is known about teachers' writing assessment practices or the conditions that foster or limit the use of such practices in their classrooms.

- *Enhancing the knowledge of best practices in writing assessment and developing evidence-based practices for minimizing or eliminating factors that bias or invalidate such assessments.*

- *Developing writing assessments that are useful to content-area teachers.* These include assessments of writing as a tool for learning and using writing to demonstrate mastery of concepts.

- *Testing curriculum-based writing measures that focus on larger units of text.* Most of these assessments concentrate on smaller units of text, such as words or sentences. A study by Espin, De La Paz, Scierka, and Roelofs (2005) reveals that using larger units of text, such as genre elements, may also be a fruitful strategy.

CONCLUSION

The findings from this report provide strong support for the use of formative writing assessment as a tool for improving students' ability to effectively convey thoughts and ideas through writing. Teacher, student, and peer assessment of writing leads to better writing. Because formative writing assessments have consequences for students (such as grades) that go beyond learning to write, it is important that they are implemented in conjunction with best practices in writing assessment.

Helping students become more skilled and confident writers has social implications beyond the classroom. Helping these young people learn to write clearly, coherently, and logically will expand their access to higher education, give them the skills needed to be successful at work, and increase the likelihood that they will actively participate as citizens of a literate society.

Improving students' writing does not rest just on improving writing assessment. Developing students who are skilled and confident writers will also require better-prepared teachers, making writing and writing instruction a priority in every teacher's classroom, and providing students with twenty-first-century writing tools. Only the combined efforts of policymakers, educators, and researchers will make this happen.

APPENDIX A: METHODOLOGY

This appendix reviews the procedures for locating the studies included in this report. It also provides information on the methodology used to compute effect sizes and conduct the meta-analyses presented in this report.

Search Strategies

Location and Selection of Studies

The strategies used for locating and selecting studies for this report were influenced by the following five factors.

First, studies had to involve students in grades one to twelve. Investigations involving writing assessment with kindergarten students, college students, and adults were excluded.

Second, the search concentrated on studies examining writing assessment. These included studies focusing on the effects of writing assessment; the reliability and validity of specific types of common writing assessments (analytic and holistic scoring for writing quality, curriculum-based assessment, and computer marking systems); and factors influencing the reliability, validity, and fairness of writing assessments. (It should be noted that reliability studies examining holistic and analytic scoring methods were limited to studies where there were more than two scorers.)

Third, the primary assessment variable of interest was writing quality (that is, students' ability to effectively convey thoughts and ideas through writing). However, studies were included that examined other writing variables or other aspects of writing assessment. These included outcome measures other than writing quality when examining the effects of progress monitoring (as writing quality was the outcome in only two studies) as well as the reliability of curriculum-based assessment and computer marking systems (these measures typically assess more discrete aspects of students' writing).

Fourth, recommendation 1 and most of the findings for recommendation 2 are based on effect sizes computed from true-experimental or quasi-experimental studies. For recommendation 2, one of the findings where effect sizes were computed included three studies where students served as their own control (i.e., students wrote a composition both by hand and on a word processor). Similarly, one study looks at the impact of computer marking systems that involved students serving as their own controls. True-experimental, quasi-experimental, or subjects-as-own-control studies in these analyses were not included if the data needed to calculate appropriate statistics for an effect size and average weighted effect size was not available.

Fifth, a search that was as broad as possible was undertaken to identify relevant studies for this review. In December 2009, electronic searches were run in multiple databases, including ERIC, PsychINFO,

ProQuest, Education Abstracts, and Dissertation Abstracts, to identify relevant studies. Descriptors included assessment, evaluation, portfolio, performance assessment, curriculum-based assessment, curriculum-based measurement, automated essay scoring, computer scoring, analytic quality, holistic quality, word processing, self-assessment, feedback, peer feedback, high-stakes assessment, state writing assessments, handwriting and writing quality, spelling and writing quality, and grammar and writing quality.

Almost 7,000 items were identified through the electronic searches. Each entry was read by the first authors of this review. If the item looked promising based on its abstract or title, it was obtained. In addition, hand searches were conducted for the following peer-reviewed journals: *Assessing Writing*, *Research in the Teaching of English*, and *Written Communication*. Once a document was obtained, the reference list was searched to identify additional promising studies. Of 512 documents collected, 136 documents were found that contained experiments that met the inclusion criteria.

Categorizing Studies According to Questions and Methods

This report was conducted to answer the following questions:

1. Does formative writing assessment enhance students' writing?
2. How can teachers improve formative writing assessment in the classroom?

Each study was read and then placed into a category based on the question it was designed to answer. If it did not provide information relevant to one of the two questions above, it was placed in an exclusion category. This process was repeated several times, resulting in multiple subcategories for each question. Categories for question 1 included the impact of feedback, self-assessment, progress monitoring, and computer marking systems. The categories for question 2 included the effects of word processing during testing, presentation effects on writing quality, knowledge of writer effects, context effects, relationship between the quality of students' writing performance across genres, and reliability of writing measures (curriculum-based, holistic quality, analytic quality). Once these subcategories were created, all studies, including the ones that were initially excluded, were reexamined to determine if they belonged in their assigned category and if other categories needed to be created. This final examination resulted in the movement of three studies. No new categories were created, however.

Meta-analysis

Calculation of Effect Sizes for Studies Included in Recommendations 1 and 2

The studies in the meta-analyses included designs where randomization did (experimental) and did not (quasi-experimental) occur. The meta-analyses also included four studies where subjects acted as their own controls. When a writing pretest measure comparable to the posttest measure was available

for either a true or quasi-experiment, an effect size (d) was computed as the difference between the treatment and control condition (i.e., $\overline{Y}_{tx} - \overline{Y}_{ctrl}$) after adjusting for pretest writing differences by subtracting the mean difference at pretest from posttest, or estimating the posttest mean-difference statistic from covariate-adjusted posttest means. This difference was then divided by the pooled standard deviation for the posttest. In a few instances, it was necessary to compute an effect size for the posttest and pretest separately, and obtain an adjusted effect size by subtracting the effect size for the pretest from the effect size for the posttest. In each of these cases, the pretest and posttest were measures of the same construct, but used different scales for measuring it. When a pretest was not available, effect sizes were calculated by subtracting the mean posttest performance of the control group from the mean posttest performance of the writing treatment group and dividing by the pooled standard deviation of the two groups. All computed effects were adjusted for small-sample size bias ($d_{adj} = d * \gamma$; $\gamma = 1 - 3/4[n_{tx} + n_{ctrl}] - 9$; Hedges, 1982).

For both experimental and quasi-experimental designs, missing standard deviations were estimated from summary statistics reported by researchers or by estimating residual sums of squares to compute a root mean squared error (RMSE) (e.g., Shadish, Robinson, and Congxiao, 1999; Smith, Glass, and Miller, 1980). For covariate or complex factorial designs, pooled standard deviations were estimated by calculating and restoring the variance explained by covariates and other "off-factors" to the study's error term and recalculating the RMSE, or pooled standard deviation, from the composite variance.

As a prelude to calculating the effect size for some comparisons, it was necessary to average performances of two or more groups in each condition. For example, some studies provided separate statistics by grade or type of writer for the treatment and control conditions. To aggregate data in each condition, the procedure recommended by Nouri and Greenberg (described in Cortina and Nouri, 2000) was applied. This procedure estimates an aggregate group or grand mean and provides a correct calculation of the variance by combining the variance within and between groups. We first calculated the aggregate treatment or control mean as an n-weighted average of subgroup means:

$$\overline{Y}_{\bullet\bullet} = \frac{1}{n_{\bullet\bullet}} \left[\sum_{j=1}^{k} \left(n_{\bullet j} \right) \left(\overline{Y}_{\bullet j} \right) \right]$$

Then, the aggregate variance was calculated by adding the n-weighted sum of squared deviations of group means from the grand mean to the sum of squared deviations within each subgroup:

$$s^2_{\bullet\bullet} = \frac{1}{n_{\bullet\bullet} - 1} \left[\sum_{j=1}^{k} n_{\bullet j} \left(\overline{Y}_{\bullet\bullet} - \overline{Y}_{\bullet j} \right)^2 + \sum_{j=1}^{k} \left(n_{\bullet j} - 1 \right) s^2_{\bullet j} \right]$$

Aggregated treatment or control means and standard deviations were used to compute an independent effect size (*d*).

Across studies, there was no single writing quality measure used by investigators. Measures of writing quality were based on examiners' judgment of the overall merit of a paper, taking into account factors such as ideation, organization, vocabulary, sentence structure, and tone. These attributes were assessed singularly (analytic scale) or together (holistic scale) on a numerical Likert-type rating scale. If a holistic score was available, we calculated the effect size with this measure. If both holistic and analytic scores were available, only the holistic score was used. If just an analytic scale was available, we first calculated an effect size for each attribute separately, and then averaged these separate effect sizes to obtain a global measure of writing quality (similar to a holistic score).

Statistical Analysis of Effect Sizes

An average weighted effect size was computed for a recommendation when there were at least four or more independent comparisons assessing the same issue. Although both Hillocks (1986) and Graham and Perin (2007) applied the same criterion, it must be recognized that small sample sizes are not very reliable, and a summary statistic is not reported with small samples and considerable variation in effect sizes. There was one exception to the rule of four. First, we did compute an average weighted effect size for the three subjects-as-own-control studies examining the relationship between word processing skills and students' performance when using a word processor versus writing by hand (recommendation 2). It should also be noted that we did not compute an average weighted effect size for context effects in recommendation 2, where only two studies were available. Instead, we simply presented the two effect sizes.

Our meta-analysis employed a weighted random-effects model. For each meta-analysis, we calculated the mean and confidence interval for the average weighted effect size. While it is best to interpret the magnitude of an effect size in relation to the distribution of other mean effect sizes in the same general area, a widely used rule of thumb is that an effect size of 0.20 is small, 0.50 is medium, and 0.80 is large.

We further conducted tests of homogeneity to determine if the various effect sizes weighted and averaged together for a specific recommendation estimated the same population effect size. When variability in effect sizes was larger than expected based on sampling error alone (i.e., the homogeneity test was statistically significant), and there were at least twelve effect sizes computed for the treatment, we examined if this excess variability could be accounted for by identifiable differences between studies (e.g., training versus no training). Using a random-effects model (Lipsey and Wilson, 2001), effect sizes were partitioned to determine if a specific study feature accounted for a significant proportion of the excess variability in effect sizes.

To avoid inflating sample size and violating the assumption of independence of data (Wolf, 1986), only one effect size for each study was used when conducting the analysis for each recommendation. Note that not all of the data from the meta-analyses are included in this document (e.g., tests of homogeneity). These can be obtained from the first author (Steve Graham, at steve.graham@vanderbilt.edu).

REFERENCES

Achieve, Inc. (2005). *Rising to the challenge: Are high school graduates prepared for college and work?* Washington, DC: Author.

ACT. (2005). *College readiness standards.* Iowa City, IA: Author. Retrieved from www.act.org

Albertson, B. (2007). Organization and development features of grade 8 and grade 10 writers: A descriptive study of Delaware student testing program (DSTP) essays. *Research in the Teaching of English, 41,* 435–465.

Andrade, H. L., Wang, X., Du, Y., and Akawi, R. L. (2009). Rubric-referenced self-assessment and self-efficacy for writing. *Journal of Education Research, 102,* 287–301.

Andrews, R., Torgerson, C., Beverton, S., Freeman, A., Locke, T., Low, G., and Zhu, D. (2006). The effects of grammar teaching on writing development. *British Educational Research Journal, 32,* 39–55.

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

Beck, S. W., and Jeffery, J. V. (2007). Genres of high-stakes writing assessments and the construct of writing competence. *Assessing Writing, 12,* 60–79.

Ben-Simon, A., and Bennett, R. E. (2007). Toward more substantively meaningful automated essay scoring. *Journal of Technology, Learning, and Assessment, 6*(1), 1–47. Retrieved from http//www.jtla.org

Berman, I. (2009). Supporting adolescent literacy achievement. Issue brief. Washington, DC: National Governors Association, 1–15.

Biancarosa, G., and Snow, C. (2004). *Reading next—A vision for action and research in middle and high school literacy: A report to Carnegie Corporation of New York.* Washington, DC: Alliance for Excellent Education.

Brown, G., Glasswell, K., and Harland, D. (2004). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing, 9,* 105–121.

Callahan, S. (1999). All done with the best of intentions: One Kentucky high school after six years of state portfolio tests. *Assessing Writing, 6,* 5–40.

Carnevale, A., and Derochers, D. (2004). *Standards for what? The economic roots of K–16 reform.* Princeton, NJ: ETS.

Coniam, D. (2009). Experimenting with a computer essay-scoring program based on ESL student writing scripts. *European Association for Computer Assisted Language Learning, 21,* 259–279.

Cortina, J. M., and Nouri, H. (2000). *Effect size for ANOVA designs* (Vol. 129). Thousand Oaks, CA: Sage Publications Inc.

Couzijn, M., and Rijlaarsdam, G. (2005). Learning to write instructive texts by reader observation and written feedback. *Learning to Write: Reader Observation, 14,* 209–240.

Dappen, L., Isernhagen, J., and Anderson, S. (2008). A statewide writing assessment model: Student proficiency and future implications. *Assessing Writing, 13,* 45–60.

DeVoss, D., Eidman-Aadahl, E., and Hicks, T. (2010). *Because digital writing matters*. San Francisco, CA: Jossey-Bass.

Espin, C., Weissenburger, J., and Benson, B. (2004). Assessing the writing performance of students in special education. *Exceptionality, 12,* 55–66.

Espin, C., De La Paz, S., Scierka, B., and Roelofs, L. (2005). The relationship between curriculum-based measures in written expression and quality and completeness of expository writing for middle school students. *Journal of Special Education, 38,* 208–217.

Gewertz, C. (2011). Common-assessment consortia add resources to plans. *Education Week, 30,* 8.

Gewertz, C., and Robelen, E. (2010). U.S. tests awaiting big shifts: Most states part of groups winning federal grants. *Education Week, 30*(3), 1, 18–19.

Gilbert, J., and Graham, S. (2010). Teaching writing to elementary students in grades 4 to 6: A national survey. *Elementary School Journal, 110,* 494–518.

Graham, S. (1982). Written composition research and practice: A unified approach. *Focus on Exceptional Children, 14,* 1–16.

Graham, S. (1990). The role of production factors in learning disabled students' compositions. *Journal of Educational Psychology, 82,* 781–791.

Graham, S. (2006). Writing. In P. Alexander and P. Winne (Eds.), *Handbook of educational psychology* (pp. 457–477). Mahway, NJ: Erlbaum.

Graham, S., and Hebert, M. (2010). *Writing to read: Evidence for how writing can improve reading.* Washington, DC: Alliance for Excellence in Education.

Graham, S., and Leone, P. (1987). Effects of emotional and behavioral disability labels, quality of writing performance, and examiner's level of expertise on the evaluation of written products. *Journal of Experimental Education, 55,* 89–94.

Graham, S., and Perin, D. (2007). *Writing next: Effective strategies to improve writing of adolescents in middle and high schools.* Washington, DC: Alliance for Excellent Education.

Graham, S., Capizzi, A., Hebert, M., and Morphy, P. (2010). *Teaching writing to middle school students: A national survey*. Submitted for publication.

Haladyna, T., and Hess, R. (1999–2000). An evaluation of conjunctive and compensatory standard-setting strategies for test decisions. *Educational Assessment, 6,* 129–153.

Hales, L., and Tokar, E. (1975). The effect of the quality of preceding responses on the grades assigned to subsequent responses to an essay question. *Journal of Educational Measurement, 12,* 115–117.

Hedges, L.V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin, 92,* 490–499.

Hillocks, G. (1986). *Research on written composition: New directions for teaching*. Urbana, IL: National Council of Teachers of English.

Hillocks, G. (2002). *The testing trap: How state writing assessments control learning*. New York, NY: Teachers College Press.

Hughes, D., Keeling, B., and Tuck, B. (1980). The influence of context position and scoring method on essay scoring. *Journal of Educational Measurement, 17,* 131–135.

James, A. (1927). The effect of handwriting on grading. *English Journal, 16,* 180–205.

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

Kirsch, I., Braun, H., Yamamoto, K., and Sum, A. (2007). *America's perfect storm: Three forces changing our nation's future.* Princeton, NJ: ETS.

Kiuhara, S., Graham, S., and Hawken, L. (2009). Teaching writing to high school students: A national survey. *Journal of Educational Psychology, 101,* 136–160.

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

Lipsey, M., and Wilson, D. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

MacArthur, C. A., Schwartz, S. S., and Graham, S. (1991). Effects of a reciprocal peer revision strategy in special education classrooms. *Learning Disabilities Research and Practice, 6,* 201–210.

Moon, T., and Hughes, K. (2002). Training and scoring issues involved in large-scale writing assessments. *Educational Measurement: Issues and Practice, 21,* 15–19.

Nagin, C. (2003). *Because writing matters: Improving students' writing in our schools.* San Francisco, CA: Jossey-Bass.

National Center for Education Statistics. (2005). *A first look at the literacy of America's adults in the 21st century.* Washington, DC: U.S. Government Printing Office.

National Commission on Writing. (2003). *The neglected R: The need for a writing revolution*. Retrieved from www.collegeboard.com

National Commission on Writing. (2004). *Writing: A ticket to work or a ticket out: A survey of business leaders.* Retrieved from www.collegeboard.com

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

National Commission on Writing. (2006). *Writing and school reform*. New York: College Board. Retrieved from www.collegeboard.com

National Institute of Child Health and Human Development. (2000). *Report of the national reading panel: Teaching students to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups.* Bethesda, MD: National Institute of Child Health and Human Development, National Institutes of Health.

Nunnally, J. (1967). *Psychometric theory.* New York, NY: McGraw-Hill.

Parke, C., Lane, S., and Stone, C. (2006). Impact of a state performance assessment program in reading and writing. *Educational Research and Evaluation, 12,* 239–269.

Popp, S., and Ryan, J. (2003). *The effect of benchmark selection on the assessed quality of writing.* Paper presented at the State and Regional Educational Research Association annual meeting, Chicago, IL.

Rogers, L., and Graham, S. (2008). A meta-analysis of single subject design writing intervention research. *Journal of Educational Psychology, 100,* 879–906.

Rosenthall, B. D. (2006). Improving elementary-age children's writing fluency: A comparison of improvement based on performance feedback frequency. (Unpublished doctoral dissertation). Syracuse University, Syracuse, NY.

Russell, M., and Plati, T. (2000). Mode of administration effects on MCAS composition performance for grades four, eight, and ten. A report of findings submitted to the Massachusetts Department of Education (World Wide Web Bulletin), Chestnut Hill, MA.

Russell, M., and Tao, W. (2004a). The influence of handwriting and computer-print on composition scores: A follow-up to Powers, Fowles, Farnum, and Ramsey. *Practical Assessment, Research and Evaluation, 9*(1), 1–17. Retrieved from http://pareonline.net/getvn.asp?v=9andn=1

Russell, M., and Tao, W. (2004b). The influence of computer-print on rater scores. *Practical Assessment Research and Evaluation, 9,* 1–17.

Salahu-Din, D., Persky, H., and Miller, J. (2008). *The nation's report card: Writing 2007* (NCES 2008–468). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Scammacca, N., Roberts, G., Vaughn, S., Edmonds, M., Wexler, J., Reutebuch, C., and Torgesen, J. (2007). *Intervention for struggling readers: A meta-analysis with implications for practice.* Portsmouth, NH: RMC Research Corp.

Schunk, D. H., and Swartz, C. W. (1993). Goals and progress feedback: Effects on self-efficacy and writing achievement. *Contemporary Educational Psychology, 18*(3), 337–354.

Shadish, W. R., Robinson, L., and Congxiao, L. (1999). *ES: A computer program for effect size calculation.* Memphis, TN: University of Memphis.

Shavelson, R., and Webb, N. (1991). *Generalizability theory: A primer.* Thousand Oaks, CA: Sage.

Smith, M. L., Glass, G.V., and Miller, T. I. (1980). *The benefits of psychotherapy.* Baltimore, MD: Johns Hopkins University Press.

Taggart, A., et al. (2001). *The national economic downturn and determining youth employment prospects: The case for a young adult job stimulus program.* Chicago, IL: Alternative School Network.

Wolf, I. (1986). *Meta-analysis: Quantitative methods for research synthesis.* Beverly Hills, CA: Sage.

APPENDIX B: STUDIES AND DATA EXAMINED IN THE REPORT

Table 1. Effect sizes for providing feedback on quality of students' writing (recommendation 1)

Students Receive Feedback from an Adult or Peer

Study	Design	Grade	Treatment	N	ES
Rosenthall, 2006	Quasi	3	Teachers provided students with feedback on writing output and spelling accuracy either once a week or three times a week	42	0.29
Couzijn and Rijlaarsdam, 2005	True-exp	9	Students observed a peer try to perform a task they wrote about and also received written feedback	54	2.37
Guastello, 2001	Quasi	4	Parents gave students feedback on their written work	167	1.05
Couzijn, 1999	True-exp	9	Students listened to a peer evaluate whether their text was or was not an argument	60	0.43
Lumbelli, Paoletti, and Frausin, 1999	True-exp	6	Students practiced revising text after receiving verbal feedback from an adult about unclear or missing information in text	28	0.83
Schunk and Swartz, 1993a (study 1)	True-exp	5	Teachers provided students with their progress in learning a writing strategy	30	0.68
Schunk and Swartz, 1993a (study 2)	True-exp	4	Teachers provided students with their progress in learning a writing strategy	20	0.83
Schunk and Swartz, 1993b (study 2)	True-exp	4	Teachers provided students with their progress in learning a writing strategy	22	1.42

Students Give Feedback to Other Students

Study	Design	Grade	Treatment	N	ES
Paquette, 2009	Quasi	2	Students were taught how to analyze the writing of an older student who was acting as their tutor	35	-0.02
Holliway and McCutchen, 2004	True-exp	5 and 9	Students gave three other students feedback about their writing	98	0.25

Students Receive and Give Feedback from and to Their Peers

Study	Design	Grade	Treatment	N	ES
Boscolo and Ascorti, 2004	Quasi	4, 6, 8	Students gave and received feedback on their writing to and from each other	122	0.93
Prater and Bermudez, 1993	True-exp	4 (ELL)	Students gave and received feedback on their writing to and from each other	46	0.15
Wise, 1992	Quasi	8	Students gave and received feedback on their writing to and from each other	88	0.65
MacArthur, Schwartz, and Graham, 1991	Quasi	4–6 (LD)	Students were taught strategies for giving and receiving feedback to and from each other	29	1.42
Olson, 1990	Quasi	6	Students gave and received feedback on their writing to and from each other	44	1.14
Benson, 1979	Quasi	7–8	Students gave and received information and reinforcing feedback about their writing to and from each other	126	0.37

Note: N = Number of students; ES = Effect size; Quasi = Quasi-experimental design; True-exp = True-experimental design; ELL = English language learners; LD = Students with learning disabilities.

Table 2. Effect sizes on writing quality for teaching students to assess their own writing (recommendation 1)

Study	Design	Grade	Treatment	N	ES
Andrade, Du, and Wang, 2008	Quasi	3–4	Students received minimal instruction on how to use a rubric to score their writing	106	0.87
Andrade and Boulay, 2003	Quasi	7–8	Students received minimal instruction on how to use a rubric to score their writing	107	0.00
Duke, 2003	True-exp	10–12	Students taught how to use a rubric to score their compositions	164	0.29
Guastello, 2001	Quasi	4	Students taught how to use a rubric to score their compositions	167	1.27
Ross, Rolheiser, and Hogboam-Gray, 1999	Quasi	4–6	Students taught how to use a rubric to score their compositions	296	0.20
Reynolds, Hill, Swassing, and Ward, 1988	Quasi	6–8	Students taught two strategies for evaluating their writing	54	0.15
Fitzgerald and Markham, 1987	True-exp	6	Students taught how to evaluate their writing	30	0.31

Note: N = Number of students; ES = Effect size; Quasi = Quasi-experimental design; True-exp = True-experimental design.

Table 3. Effect sizes for teachers monitoring students' writing performance (recommendation 1)

Study	Design	Grade	Type	Areas Assessed	Outcome	N	ES
Crehan and Curfman, 2003	Quasi	8	ANALY	I, O, V, C	Analytic quality	269 Ss	0.35
Jewell, 2003	True-exp	3, 5, 8	CBM	CWS, WSC, TWW	Analytic quality	257 Ss	0.12
Vellella, 1996	Quasi	2	CBM	SP	SP–norm-ref	77 Ss	0.12
Fuchs, Fuchs, and Hamlett, 1989	True-exp	ELEM	CBM	SP	SP–norm-ref	27 classes	0.26
Fuchs, Fuchs, Hamlett, and Allinder, 1991a	True-exp	ELEM	CBM	SP	SP–norm-ref	30 classes	0.26
Fuchs, Fuchs, Hamlett, and Allinder, 1991b	True-exp	ELEM	CBM	SP	SP–norm-ref	30 classes	0.26
Fuchs, Fuchs, and Hamlett, 1989	True-exp	ELEM	CBM	SP	SP–norm-ref	27 classes	0.26

Note: Type = Type of assessment (scoring writing using a rubric with scores for specific writing traits or curriculum-based assessment procedures); Quasi = Quasi-experimental design; True-exp = True-experimental design; ANALY = Analytic quality measure; I = Ideas; O = Organization; V = Voice; C = Conventions; CWS = Correct word sequence; WSC = Words spelled correctly; TWW = Total words written; SP = Spelling; Ss = Students; CBM = Curriculum-based measurement; Norm-ref = Norm-referenced test; ELEM = Elementary students. Analytic quality involves summing separate scores for prominent features of writing, such as ideation, organization, voice, vocabulary, sentence variety, and conventions.

Table 4. Effect sizes for the effect of word processing on the quality of students' writing when taking writing tests (recommendation 2)

Studies Comparing Word Proccessing to Writing by Hand (No Attempt Made to Determine Effects of Word Processing Experience)

Study	Design	Grade	Student	N	ES
Russell and Plati, 2000 (study 1)	Quasi	4	FR	144	0.34
Russell and Plati, 2000 (study 2)	Quasi	8	FR	144	0.65
Russell and Plati, 2000 (study 3)	Quasi	8	FR	84	0.77
Russell and Plati, 2000 (study 4)	Quasi	10	FR	145	0.54
Russell, 1999	True-exp	MS	FR*	117	0.04
Russell and Haney, 1997	Quasi	6–8	FR	86	0.89
Wolfe, Bolton, Feltovich, and Welch, 1993	Quasi	10	FR	155	0.62

Studies Comparing More and Less Experienced Word Processors (Students Wrote by Hand Versus Word Processing)

Study	Design	Grade	Student	N	ES
Burke and Cizek, 2006	Ss own control	6	Skilled Less skilled	158	Skilled = -0.21 Less = 0.35
Wolfe, Bolton, Feltovich, and Niday, 1996	Ss own control	HS	High exper Middle exper Low exper	60	High = -0.36 Med = 0.00 Low = 1.12
Wolfe, Bolton, Feltovich, and Bangert, 1996	Ss own control	10	High exper Low exper	406	High = 0.00 Low = 0.37

Note: N = Number of students; ES = Effect size; Quasi = Quasi-experimental design; True-exp = True-experimental design; Ss own control = Students as own controls; FR = Full range of abilities; FR* = Typical students with little word processing experience; MS = Middle school; HS = High school; Exper = Experience.

Table 5. Effect sizes for presentation effect on writing quality (recommendation 2)

More Legible Versus Less Legible Writing

Study	Grade	Raters	N	ES
Klein and Taub, 2005	6	T	53 raters	-1.10
Soloff, 1973	11	T	32 raters	-0.94
Marshall and Powers, 1969	12	TIT	70 raters	-0.38
Sheppard, 1929 (study 1)	8	T	450 raters	-1.10
Sheppard, 1929 (study 2)	8	T	450 raters	-1.30

Typed Versus Handwritten Text

Study	Grade	Raters	N	ES
Russell and Tao, 2004a (study 1)	8	T	60 Ss	-0.47
Russell and Tao, 2004b	4	T	52 Ss	-0.64
	8	T	60 Ss	-0.83
	10	T	60 Ss	-0.55
Wolfe, Bolton, Feltovich, and Welch, 1993	10	T (one-half of raters)	157 Ss	-0.27

Note: Papers that were typed, had spelling errors, were less legible, or had grammar errors yielded lower scores than papers with the same content that were handwritten, had no spelling errors, were more legible, or had no grammatical errors, respectively; N = Number of student papers scored (Ss) or number of raters; ES = Effect size; T = Teachers; TIT = Teachers in training.

Table 6. Effect sizes for knowledge of the writer effect on writing quality (recommendation 2)

Study	Grade	Raters	N (Raters)	Status of Writer	ES
Peterson, Childs, and Kennedy, 2004	6	T	108	Girls vs. boys	-0.38
Davidson, Hoekema, and Howell, 2000	8	T	144	Ethnic minority vs. majority	-0.73
Howell, Bigelow, Moore, and Evoy, 1993	8	T	200	Ethnic minority vs. majority	-0.83
Graham and Dwyer, 1987	4	TIT	22	Learning disability vs. no disability	-1.15
Graham and Leone, 1987	8	TIT	88	Behavioral disability vs. no disability	-0.23

Note: A negative effect means that knowledge of writer characteristics (e.g., knowing the writer was a girl) had a negative effect on rating of writing quality; ES = Effect size; T = Teachers; TIT = Teachers in training.

Table 7. Effect sizes for context effect on writing quality (recommendation 2)

Study	Grade (Age)	Raters	N	ES
Hughes, Keeling, and Tuck, 1980	(13–14)	TIT	212	-0.72
Hales and Tokar, 1975	5–6	TIT	128	-0.31

Note: A negative effect means an essay of average quality was scored lower if the essays scored before it were of high quality; N = Number of raters; ES = Effect size; TIT = Teachers in training.

Table 8. Differences in students' performance on different writing tasks (recommendation 2)

Study	Tasks	Grade	Quality Measure	Statistical Differences
Hebert, Graham, and Harris, 2010	N, P, I	2–3	Holistic	N and I > P (grade 2), P > N and I
Popp, Ryan, Thompson, and Behrens, 2003	N, RESP, P	5 and 8	Analytic	N > P (grade 8)
Engelhard, Gordon, Gabrielson, and Walker, 1994	N, D, E	8	Analytic	All three modes differed statistically on topic development
Gearhart, Herman, Baker, and Whitaker, 1992	N, SUM	3–4	Holistic	N > SUM
Engelhard, Gordon, and Gabrielson, 1991	N, D, E	8	Analytic	All three modes differed statistically for C/O and ST
Carlman, 1985	EXP, TRANS	12	Holistic	EXP > TRANS
Veal and Tillman, 1971	P, D, E, N	2, 4, 6	Holistic	N > P, D > P, E > N (grade 4), D, E, N > P, D, E, E > N (grade 6)

Note: P = Persuasive; D = Descriptive; E = Expository; N = Narrative; RESP = Response to literature; SUM = Summary; EXP = Expressive; TRANS = Transactional; I = Informative. Holistic quality involves assigning a single score to a piece of writing, taking into account multiple aspects of writing; analytic scoring involves summing separate scores for prominent features of writing, such as ideation, organization, voice, vocabulary, sentence variety, and conventions.

Table 9. Correlations between students' performance on different writing tasks (recommendation 2)

Study	Task	Grade (Age)	Quality Measure	Correlations
Lane et al., in press	S, P	2	Holistic	.24–.55
Hebert, Graham, and Harris, 2010	N, PN, P, I	2–3	Holistic	.44–.60
Popp, Ryan, Thompson, and Behrens, 2003	RESP, N, P	5 and 8	Analytic	.56 (grade 5) .59 (grade 8)
Hunter, Jones, and Randhawa, 1996	NR	5, 8, 11	Analytic	.27–.34
Purves, 1992	N, P, REFL, EXP	NR	Holistic	.32 (U.S.) .2–.61 (across twelve countries)
Hogan and Mishler, 1980	N	3 and 8	Holistic	.71 (grade 3) .77 (grade 8)
Lehmann, 1990	LET, N, P, A, RHET	11	Holistic	.10–.46
Swartz and Whitney, 1985	P, D, E	12	Holistic	.54–.65
Moss, Cole, and Khampalikit, 1982	I, PUR	4, 7, 10	Holistic	.41 (grade 4) .50 (grade 7) .46 (grade 10)
Quellmalz, Capell, and Chou, 1982	N and E	11–12	Holistic, analytic	.43 (holistic) .23 (analytic)
Finlayson, 1951	E	(12)	Holistic	.69

Note: S = Story; N = Narrative; PN = Personal narrative; P = Persuasive; I = Informational; RESP = Response to literature; REFL = Reflection; EXP = Expressive; NR = Not reported; LET = Letter; A = Advise; RHET = Rhetorical; D = Descriptive; E = Expository; PUR = A letter that serves a specific purpose. Holistic quality involves assigning a single score to a piece of writing, taking into account multiple aspects of writing. Analytic scoring involves summing separate scores for prominent features of writing, such as ideation, organization, voice, vocabulary, sentence variety, and conventions.

Table 10. Progress monitoring curriculum-based assessments defined (recommendation 2)

Measure	Definition
Number of correct word sequences (CWS)	The number of combinations of two adjacent, correctly spelled words that are syntactically and semantically appropriate given the context of the sentence, according to a native speaker of the English language. Correct meanings, tenses, number agreement (singular or plural), noun-verb correspondences, punctuation, capitalization, and spelling are all taken into account.
Percentage of correct word sequences (%CWS)	The ratio of the number of correct word sequences to the total number of word sequences
Correct minus incorrect word sequences (CMIWS)	The number of incorrect word sequences subtracted from the total number of correct word sequences.
Mean length of correct word sequences (MLCWS)	The number of unbroken strings of correct word sequences divided by the total number of correct word sequences.

Number of sentences (#Sent)	Any series of words separated from another series of words by a period, question mark, or exclamation point. The series of words must include a recognizable subject and verb, but does not need to contain correct capitalization or punctuation.
Total written words (TWW)	The total number of words written in the allotted test time. Spelling, grammar, and content are not taken into consideration. Numerals and symbols are not counted. However, some studies did not require words to be legible to be counted as a word.
Words spelled correctly (WSC)	A count of the total number of legible words that can stand alone as a correctly spelled word in the English language. Context and grammar are not taken into account. For example, if an incorrect homonym is used in context (e.g., "their" instead of "there"), the word is still counted as a correctly spelled word.
Percentage of words spelled correctly (%WSC)	The ratio of the number of words spelled correctly to the total number of words written.
Correct letter sequences (CLS)	Any two adjacent letters that are in the correct sequence according to the correct spelling of the word in the English language. (In some studies, researchers count the first letter and the last letter as "stand-alone" correct letter sequences. For example, in these instances, the total number of correct letter sequences in the word "cat" would be four: "c," "ca," "at," and "t.")

Note: Only writing curriculum-based measures tested in four or more studies are included in this table.

Table 11. Reliability of curriculum-based assessments (recommendation 2)

Correct Word Sequences

Study	Grade Level	Genre	Time (Min.)	Reliability	
				% Consensus	*R* Consistency
Amato and Watkins, 2011	8	N	3		> .94
Coker and Ritchey, 2010	1	Write sent.	6	91%	
McMaster et al., 2009 (study 1)	1	Copy sent.	5	> 92%	
McMaster et al., 2009 (study 1)	1	N	5	> 92%	
McMaster et al., 2009 (study 2)	1	Write sent.– pict. and word prompts	5	> 92%	
McMaster et al., 2009 (study 2)	1	N (pict. prompt)	5	> 92%	
McMaster et al., 2009 (study 2)	1	N (pict. prompt)	5	> 92%	
Ternezzi, 2009	4	N	NR	94%	
Amato, 2008	8				.99
Espin et al., 2008	10	N	10	96%	
Crawford et al., 2006	3, 5, 8, 10	N	NR		> .88
Gansle et al., 2006	1–5	N	4	94%	
Espin et al., 2005	7–8	P and E	35	97–98%	
Jewell and Malecki, 2005	2, 4, 6	N	4		> .98

Weissenburger and Espin, 2005	4, 8, 10	N	10.5		NR
Gansle et al., 2004	3–4	N	3	93%	
Jewell, 2003	3, 5, 8	N	4		.97–.99
Lembke et al., 2003	4	Copy sent.	3		NS
Lembke et al., 2003	5	Dict. sent.	3		NS
Malecki and Jewell, 2003	1–8	N	10		.99
Gansle et al., 2002	3–4	N	3	86%	
Bath, 2000	4–6	N	3.5		.99
Espin et al., 1999	10	N	3.5	97%	
Nolet and McLaughlin, 1997 (Written Expression)	5	N	15	90–98%	
Nolet and McLaughlin, 1997 (Written Retell)	5	N	15	90–98%	
Hubbard, 1996	3	N	4	82%	.99
Watkinson and Lee, 1992	6–8	N	7		.95
Parker et al., 1991a	2–8, 11	N	7		.87
Parker et al., 1991b	6–8	N	6		.87
Tindal and Parker, 1991	3–5	N	10		.92
Tindal and Parker, 1989a	6–8	N	6.5		.87
Videen et al., 1982	3–6	N	5	86–91%	

Percentage of Correct Word Sequences

Study	Grade Level	Genre	Time (Min.)	Interrater Agreement	
				%	*R*
Amato and Watkins, 2011	8	N	3		> .94
Du, 2009	4, 6–8	N	10	97%	
Ternezzi, 2009	4	N	NR	95%	
Amato, 2008	8				.95
Jewell and Malecki, 2005	2, 4, 6	N	4		> .98
Jewell, 2003	3, 5, 8	N	4		.97–.99
Malecki and Jewell, 2003	1–8	N	10		NS
Nolet and McLaughlin, 1997 (Written Expression)	5	N	15		.95–1.0
Nolet and McLaughlin, 1997 (Written Retell)	5	N	15		.95–1.0

Watkinson and Lee, 1992	6–8	N	7		.82
Parker et al., 1991a	2–8, 11	N	7		.87
Tindal and Parker, 1989a	6–8	N	6.5		.87

Correct Minus Incorrect Word Sequences

Study	Grade Level	Genre	Time (Min.)	Interrater Agreement	
				%	*R*
Amato and Watkins, 2011	8	N	3		> .94
Du, 2009	4–8	N	10	98%	
McMaster et al., 2009 (study 1)	1	Copy words	5	> 92%	
McMaster et al., 2009 (study 1)	1	Copy sent.	5	> 92%	
McMaster et al., 2009 (study 1)	1	N	5	> 92%	
McMaster et al., 2009 (study 2)	1	Write sent.–pict. and word prompt	5	> 92%	
McMaster et al., 2009 (study 2)	1	N–pict. prompt	5	> 92%	
McMaster et al., 2009 (study 2)	1	N–photo prompt	5	> 92%	
Ternezzi, 2009	4	N	NR	86%	
Amato, 2008	8				.98
Espin et al., 2008	10	N	10	92%	
Espin et al., 2005	7, 8	P and E	35	90–91%	
Jewell and Malecki, 2005	2, 4, 6	N	4		> .98
Weissenburger and Espin, 2005	4, 8, 10	N	10.5		NR
Lembke et al., 2003	4	Copy sent.	3		NS
Lembke et al., 2003	5	Dict. sent.	3		NS
Espin et al., 2000	7, 8	D and N	5.5	88–92%	

Mean Length of Correct Word Sequences

Study	Grade Level	Genre	Time (Min.)	Interrater Agreement	
				%	*R*
Espin et al., 2000	7, 8	D and N	5.5	86–90%	
Espin et al., 1999	10	N	3.5	99%	
Parker et al., 1991b	6–8	N	6		.83
Tindal and Parker, 1989a	6–8	N	6.5		.83

Number of Sentences

Study	Grade Level	Genre	Time (Min.)	Interrater Agreement	
				%	*R*
Amato and Watkins, 2011	8	N	3		> .94
Ternezzi, 2009	4	N	NR	96%	
Amato, 2008	8	NR			.96
Gansle et al., 2002	3, 4	N	3	76%	
Espin et al., 2000	7, 8	D	5.5		1.0
Espin et al., 2000	7, 8	N	5.5		1.0
Espin et al., 1999	10	N	3.5	100%	

Total Written Words

Study	Grade Level	Genre	Time (Min.)	Interrater Agreement	
				%	*R*
Amato and Watkins, 2011	8	N	3		> .94
Coker and Ritchey, 2010	1	Sent.	6		.98
McMaster et al., 2009 (study 1)	1	Words	5	> 92%	
McMaster et al., 2009 (study 1)	1	Copy sent.	5	> 92%	
McMaster et al., 2009 (study 1)	1	N	5	> 92%	
McMaster et al., 2009 (study 2)	1	Words; letter prompt	5	> 92%	
McMaster et al., 2009 (study 2)	1	Sent.; pict. and word prompt	5	> 92%	
McMaster et al., 2009 (study 2)	1	N– pict. prompt	5	> 92%	
McMaster et al., 2009 (study 2)	1	N– photo prompt	5	> 92%	
Ternezzi, 2009	4	N	NR	99%	
Amato, 2008	8	NR	NR		.99
Espin et al., 2008	10	N	10	100%	
Crawford et al., 2006	3, 5, 8, 10	Sent.	NR		>.88
Crawford et al., 2006	3, 5, 8, 10	N	NR		>.88
Gansle et al., 2006	1–5	N	4	98%	
Jewell and Malecki, 2005	2, 4, 6	N	4		> .98
Gansle et al., 2004	3, 4	N	3	99%	

Jewell, 2003	3, 5, 8	N	4		.97–.99
Lembke et al., 2003	2	Copy words	2		NS
Lembke et al., 2003	2	Dict. words	3		NS
Lembke et al., 2003	2	Copy sent.	3		NS
Lembke et al., 2003	2	Dict. sent.	3		NS
Malecki and Jewell, 2003	1–8	N	10		.99
Gansle et al., 2002	3, 4	N	3	96%	
Bath, 2000	4–6	N	3.5		1.0
Espin et al., 2000	7, 8	D and N	5.5	100%	
Espin et al., 1999	10	N	3.5	100%	
Hedeker, 1997	1–7	N	3		NS
Hubbard, 1996	3	N	4	91%	.99
Watkinson and Lee, 1992	6–8	N	7		.99
Parker et al., 1991a	2–8, 11	N	7		.99
Parker et al., 1991b	6, 7, 8	N	6		.99
Tindal and Parker, 1991	3, 4, 5	N	10		.99
Tindal and Parker, 1989a	6–8	N	6.5		.99
Tindal and Parker, 1989b	6, 8, 11	SUM	15		NS
Deno et al., 1982	3–6	N	NS		NS

Words Spelled Correctly

Study	Grade Level	Genre	Time (Min.)	Interrater Agreement	
				%	*R*
Amato and Watkins, 2011	8	N	3		> .94
Coker and Ritchey, 2010	1	Sent.	6	92%	
McMaster et al., 2009 (study 1)	1	Copy words	5	> 92%	
McMaster et al., 2009 (study 1)	1	Copy sent.	5	> 92%	
McMaster et al., 2009 (study 1)	1	N	5	> 92%	
McMaster et al., 2009 (study 2)	1	Write words– letter prompt	5	> 92%	
McMaster et al., 2009 (study 2)	1	Write sent.–pict. and word prompt	5	> 92%	
McMaster et al., 2009 (study 2)	1	N– pict. prompt	5	> 92%	

McMaster et al., 2009 (study 2)	1	N– photo prompt	5	> 92%	
Ternezzi, 2009	4	N	NR	98%	
Amato, 2008	8				.99
Espin et al., 2008	10	N	10	99%	
Gansle et al., 2006	1–5	N	4	97%	
Jewell and Malecki, 2005	2, 4, 6	N	4		> .98
Jewell, 2003	3, 5, 8	N	4		.97–.99
Lembke et al., 2003	2	Copy words	2		NS
Lembke et al., 2003	2	Dict. words	3		NS
Lembke et al., 2003	2	Copy sent.	3		NS
Lembke et al., 2003	2	Dict. sent.	3		NS
Malecki and Jewell, 2003	1–8	N	10		.99
Gansle et al., 2002	3, 4	N	3	95%	
Bath, 2000	4–6	N	3.5		.99
Espin et al., 2000	7, 8	D	5.5		1.0
Espin et al., 2000	7, 8	N	5.5		1.0
Espin et al., 1999	10	N	3.5	99%	
Englebert-Johnson, 1997	3–6	N	4		NS
Hubbard, 1996	3	N	4	82%	.99
Watkinson and Lee, 1992	6–8	N	7		.96
Parker et al., 1991a	2–8, 11	N	7		.98
Parker et al., 1991b	6–8	N	6		.98
Tindal and Parker, 1991	3–5	N	10		.97
Tindal and Parker, 1989a	6–8	N	6.5		.98
Deno et al., 1982	3–6	N	NS		NS

Percentage of Words Spelled Correctly

Study	Grade Level	Genre	Time (Min.)	Interrater Agreement	
				%	*R*
Amato and Watkins, 2011	8	N	3		> .94
Amato, 2008	8	NR			.96
Jewell and Malecki, 2005	2, 4, 6	N	4		> .98
Jewell, 2003	3, 5, 8	N	4		NS
Malecki and Jewell, 2003	1–8	N	10		NS
Watkinson and Lee, 1992	6–8	N	7		.80
Parker et al., 1991a	2–8, 11	N	7		.98
Parker et al., 1991b	6–8	N	6		.89
Tindal and Parker, 1989a	6–8	N	6.5		.98

Correct Letter Sequences

Study	Grade Level	Genre	Time (Min.)	Interrater Agreement	
				%	*R*
McMaster et al., 2009 (study 1)	1	Copy words	5	> 92%	
McMaster et al., 2009 (study 1)	1	Copy sent.	5	> 92%	
McMaster et al., 2009 (study 1)	1	N	5	> 92%	
McMaster et al., 2009 (study 2)	1	Copy words	5	> 92%	
McMaster et al., 2009 (study 2)	1	Write sent.– pict. and word prompt	5	> 92%	
McMaster et al., 2009 (study 2)	1	N–pict. prompt	5	> 92%	
McMaster et al., 2009 (study 2)	1	N–photo prompt	5	> 92%	
Crawford et al., 2006	3, 5, 8, 10	Dict. words	NR		> .88
Crawford et al., 2006	3, 5, 8, 10	Dict. sent.	NR		> .88
Crawford et al., 2006	3, 5, 8, 10	Write sent.	NR		> .88
Lembke et al., 2003	2	Copy words	2		NS
Lembke et al., 2003	3	Dict. words	3		NS
Fergusson and Fuchs, 1991	SPED M = Grade 5.0	Dict. words	3–3.33	T: 93% C: 99%	

Angermeyer, 1988	2, 5	Dict. words	1.67		.92–.95
Deno et al., 1982	3–6	N	NS		NS

Note: N = Narrative; D = Descriptive; P = Persuasive; E = Explanatory; SUM = Summary; C = Computer; T = Teacher; Dict. = Dictated; Sent. = Sentences; NS = Not specified; WST = Word spelling test; LA = Language arts; NS = Not specified; LD = Learning disabled; ELL = English language learners; SPED = Special education students; M = Mean.

Table 12. Studies examining the reliability and validity of holistic writing quality measures (recommendation 2)

High-Stakes Tests

Study	Task	Range	GR (Age)	Raters	Trained	Interrater Agreement	
						% Consensus	*R* Consistency
Sevigny, Savard, and Beaudoin, 2009	E	5	(13–16)	NR	Y	63–64%	
Hunter, Jones, and Randhawa, 1996	NS	5	5, 8, 11	NR	Y		.82
De Ayala, Dodd, and Koch, 1991	N, EXP	4 or 8	HS	NR	Y	76%	
Lehmann, 1990	LET, N, P, E, PAR	5	11	T	Y	73%	.84
Swartz and Whitney, 1985	P, D, E	6	12	EXP	Y		.82–.87
Moss, Cole, and Khampalikit, 1982	I, PUR	4	4, 7, 10	NR	NR		.86–.94
Stewart and Grobe, 1979	I, P, E	4	5, 8, 11	T	Y		.90

Typical Writing Assessments

Study	Task	Range	GR (Age)	Raters	Trained	Interrater Agreement	
						% Consensus	*R* Consistency
Herman, Gearhart, and Baker, 1993	N	6	3–4	T–EXP	Y		.83–.85
Shohamy, Gordon, and Kraemer, 1992	P, EXP	5	12	EXP; NO EXP	TR; NO TR		.87–.93
Carlman, 1985	EXP, TRANS	6	12	T	Y		.79
Blok, 1985	E	10	NR	T	N		.80 (intra)
Veal and Hudson, 1983	NR	NR	10	NR	NR		.69–.76
Page and Paulus, 1968	E	NR	HS	EXP	NR		.43–.59
Hogan and Mishler, 1980	N	8	3 and 8	T	Y		.84–.96
Veal and Tillman, 1971	P, D, E, N	7	2, 4, 6	NR	NR		.58–.91

Finlayson, 1951	E	20	(12)	T–EXP	N		.74 (inter) .81 (intra)
Wiseman, 1949	SC	20	NR	NR	NR		.62–.79 (inter) .60–.85 (intra)

Portfolio

Study	Task	Range	GR (Age)	Raters	Trained	Interrater Agreement	
						% Consensus	*R* Consistency
Novak, Herman, and Gearhart, 1996 (WWYY scale)	N	6	2–6	EXP	TR	25%	.69
Novak, Herman, and Gearhart, 1996 (alternative scale)	N	6	2–6	EXP	TR	16%	.45
Herman, Gearhart, and Baker, 1993	N, SUM	6	1, 3, 4	EXP–T	Y		.41–.94

Note: E = Expository; I = Informational; INT = Interpretation; NR = Not reported; N = Narrative; EXPRES = Expressive; LET = Letter; P = Persuasive; PAR = Paragraph; D = Description; PUR = A letter that serves a specific purpose; SC = Student choice; TRANS = Transactional; SUM = Summary; NR = Not reported; Y = Yes; NS = Not specified; EXP = Experienced; T = Teachers; TR = Trained; TIT = Teachers in training; HS = High school; ELL = English language learners.

Table 13. Studies examining the reliability and validity of analytic writing quality measures (recommendation 2)

High-Stakes Tests

Study	Task	Range	Grade (Age)	Skills Scored	Trained	Interrater Agreement	
						% Consensus	*R* Consistency
Crawford and Smolkowski, 2008	N	10	5, 8	C/O, ST, L	T (TR)	49%	.61–.96
Crawford, Helwig, and Tindal, 2004	N, E, P, I	6	5, 8	I, O, SF, C	T (TR)	49%	.64–.79
Haladyna and Hess, 1999–2000	EXP, I, N, P	6	8, 10	I/O, O, V, VOC, SF, C	T–TR		.61 total, .50 each skill
Hollenbeck, Tindal, and Almond, 1999	SC	8	6	I, O, V, W, SF, C	T (TR)	47%	.57–.97
Kuhlemeier and van den Bergh, 1997	LET	5	9	I, ST, O, E	T (TR)		.83–.99
Gearhart, Herman, Novak, and Wolf, 1995	N	6	1–6	F/O, D, C	T–EXP (TR)	28–37%	.60–.63
Gabrielson, Gordon, and Engelhard, 1995	P	4	11	C/O, T, SF, C	NR (TR)		.87
Engelhard, Gordon, Gabrielson, and Walker, 1994	N, D, E	4	8	C/O, ST, SF, U, C	T and other (TR)		.82
Engelhard, Gordon, and Gabrielson, 1991	N, D, E	4	8	C/O, ST, SF, U, C	NR (TR)		.82

Typical Writing Assessment

Study	Task	Range	Grade (Age)	Skills Scored	Trained	Interrater Agreement	
						% Consensus	*R* Consistency
Burgin and Hughes, 2009	D	4	3, 4	I, ST, SF, U, C, SPEL	T (TR)	43–67%	.32–.57
Beyreli and Ari, 2009	N	4	6, 7	F, C, VOC, SF, PAR, NAR, TI, INTR, STY, CON	T–EXP (TR)		.68–.83
Mott, Etsler, and Drumgold, 2003	N	6	2, 3	TH, CH, PL, SET, COM	T–EXP (TR)	67–78%	.49–.59
Moon and Hughes, 2002	NS	4	6	COMP, ST, SF, U, C	EXP–TR	51–72%	
Gearhart, Herman, Novak, and Wolf, 1995	N	6	1–6	TH, CH, PL, SET, COM	T–EXP (TR)	39–44%	.48–.66
DiStefano and Killion, 1984	E	5	4–6	O, ST, F, SF, C, SP, P	T (TR)		.85–.87
Quellmalz, Capell, and Chou, 1982	N and E	4	11–12	GI, F, O, SUP, C	EXP (TR)		.61–.83
Page and Paulus, 1968	E	5	HS	C, O, ST, C, CREAT	EXP		.50

Portfolio

Study	Task	Range	Grade (Age)	Skills Scored	Trained	Interrater Agreement	
						% Consensus	*R* Consistency
Tezci and Dikici, 2006	N	4	(14–15)	SUBJ, CH, SET, INTRIG	T and others		.56–.86
Underwood and Murphy, 1998	Open	13	MS	PROC, CONSTIT, KNOWL, REFLEC	T	22–33%	.75–.89
LeMahieu, Gitomer, and Eresh, 1995	Open	6	6–12	ACCOMP, PROC, GROWTH	T–TR	46–57%	.74–.84
Koretz, Stecher, Klein, and McCaffrey, 1994	Open	4	4, 8	PURP, O, I, V, U	T	44–48%	.49–.63

Note: N = Narrative; E = Expository; P = Persuasive; I = Informational; D = Descriptive; EXPRES = Expressive; INST = Instructional; LET = Letter; OPEN = Many possible writing tasks; SC = Student choice; C/O = Content/organization; ST = Style; L = Language use; I = Ideation; O = Organization; SF = Sentence fluency; C = Conventions; A = Audience; PUR = Purpose; V = Voice; VOC = Vocabulary; I/O = Idea/organization; E = Elaboration; F/O = Focus/organization; FOR = Form; DET = Detail; SPEL = Spelling; PAR = Paragraph; NAR = Narration; TI = Title; INTR = Introduction; STY = Story; CON = Conclusion; TH = Theme; CH = Characters; PL = Place; SET = Setting; COM = Communication; GCONT = General content; TEXT = Text detail; PK = Prior knowledge; PRINC = Principles/context; MIS = Misconceptions; ARG = Argumentation; F = Focus; GI = General information; SUP = Support; CREAT = Creativity; HW = Handwriting; ORIG = Originality; SUBJ = Subject; ESTH = Esthetics; PROC = Control of process; CONSTIT = Consistency and challenge; KNOWL = Knowledge and work products; REFLEC = Reflection; ACCOMP = Accomplishment as writer; GROWTH = Growth as writer; PROC = Processes and resources as writer; U = Usage; T = Teachers; TR = Trained; EXP = Experienced; LAY = Layperson; NR = Not reported; NS = Not specified.

Table 14. Studies supporting methods for improving reliability for scoring writing quality (recommendation 2)

Study	Methods for Improving Reliability for Scoring Writing Quality
Shohamy, Gordon, and Kraemer, 1992	Providing training on how to score compositions
Burgin and Hughes, 2009; Coffman, 1966; Gearhart, Herman, Novak, and Wolf, 1995; Godshalk, Swineford, and Coffman, 1966; Swartz et al., 1999	Having multiple teachers score each paper
Johnson, Penny, Gordon, Shumate, and Fisher, 2005	Having them discuss and resolve differences in their scores
Burgin and Hughes, 2009; Finlayson, 1951; Godshalk, Swineford, and Coffman, 1966; Lehmann, 1990	Basing students' writing score on multiple writing tasks
De Ayala, Dodd, and Koch, 1991; Godshalk, Swineford, and Coffman, 1966.	Increasing the range of scores on the writing test
Hwang, 1930; Kan, 2007	Providing teachers with benchmarks (descriptions or examples) for each point on the scale
Penny, Johnson, and Gordon, 2000a, 2000b; Johnson, Penny, Fisher, and Kuhs, 2003	Applying a two-step scoring process where the teacher matches the composition to the closest benchmark, and then scores it again if it does not match this benchmark perfectly by adding a plus or minus to the first score

Table 15. Effects of computer-based feedback on the quality/content of students' writing

Study	Design	Grade	Automated Scoring Program	N	ES
Wade-Stein and Kintsch, 2004 (also reported in Steinhart, 2001 [study 3])	Ss as own control	6	Students wrote and revised summaries based on feedback they received from LSA software (Summary Street)—conditions were counterbalanced	52	0.60
Caccamise, Franzke, Eckoff, Kintsch, and Kintsch, 2007 (study 1)	Quasi	7–9	Students wrote summaries and revised them based on feedback from LSA software (Summary Street)	243	0.38
Franzke, Kintsch, Caccamise, Johnson, and Dooley, 2005 (also reported in Caccamise, Franzke, Eckoff, Kintsch, and Kintsch, 2007 [study 2])	Exp	8	Students wrote summaries and revised them based on feedback from LSA software (Summary Street)	111	0.35
Shermis, Burstein, and Bliss (2004)	Quasi	10	Students wrote responses for prompts in various genres and revised them based on feedback from regression based scoring software (e-rater)	835	-0.07

Note: N = Number of students; ES = Effect size; Quasi = Quasi-experimental design; Ss = Students; Exp = True-experimental design; LSA = Latent semantic analysis.

STUDIES PRESENTED AND REVIEWED IN APPENDIX B

Amato, J. (2008). Identifying CBM writing indices for eighth grade students. (Unpublished doctoral dissertation). Pennsylvania State University, State College, PA.

Amato, J. M., and Watkins, M. W. (2011). The predictive validity of CBM writing indices for eighth-grade students. *Journal of Special Education, 44,* 195–204.

Andrade, H. G., and Boulay, B. (2003). Role of rubric-referenced self assessment in learning to write. *Journal of Educational Research, 97,* 21–34.

Andrade, H. L., Du, Y., and Wang, X. (2008). Putting rubrics to the test: The effect of a model, criteria generation, and rubric-referenced self-assessment on elementary school students' writing. *Educational Measurement: Issues and Practice, 27,* 3–13.

Angermeyer, J. M. (1988). Some relations between the Wechsler Intelligence Scale for Children-Revides and the changes in scores on weekly curriculum-based measures of spelling and mathematics. (Unpublished doctoral dissertation). University of Minnesota, Minneapolis, MN.

Bath, L. M. (2000). Curriculum based measurement of written language: An alternative to traditional assessment procedures. (Unpublished master's thesis). California State University, Fresno, Fresno, CA.

Benson, N. L. (1979). The effects of peer feedback during the writing process on writing performance, revision behavior, and attitude toward writing. (Unpublished master's thesis). University of Colorado, Boulder, CO.

Beyreli, L., and Ari, G. (2009). The use of analytic rubric in the assessment of writing performance-inter-rater concordance study. *Educational Sciences: Theory and Practice, 9,* 105–125.

Blok, H. (1985). Estimating the reliability, validity, and invalidity of essay ratings. *Journal of Educational Measurement, 22,* 41–52.

Boscolo, P., and Ascorti, K. (2004). Effects of collaborative revision on children's ability to write understandable narrative text. In L. Allal, L. Chanquoy, and P. Largy (Eds.), *Revision: Cognitive and instructional processes* (pp. 157–170). Boston, MA: Kluwer Academic Pub.

Burgin, J., and Hughes, G. D. (2009). Credibly assessing reading and writing abilities for both elementary student and program assessment. *Assessing Writing, 14,* 25–37.

Burke, J. N., and Cizek, G. J. (2006). Effects of composition mode and self-perceived computer skills on essay scores of sixth graders. *Assessing Writing, 11,* 148–166.

Caccamise, D., Franzke, M., Eckhoff, A., Kintsch, E., and Kintsch, W. (2007). Guided practice in technology-based summary writing. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theory, interventions, and technologies.* Mahwah, NJ: Erlbaum.

Carlman, N. (1985). *Variations in the writing performance of grade 12 students: Differences by mode and topic.* ERIC Document Reproduction Services No. ED269766.

Coffman, W. E. (1966). On the validity of essay tests of achievement. *Journal of Educational Measurement, 3,* 151–156.

Coker, D. L., and Ritchey, K. D. (2010). Curriculum-based measurement of writing in kindergarten and first grade: An investigation of production and qualitative scores. *Exceptional Children, 76,* 175–193.

Couzijn, M. (1999). Learning to write by observation of writing and reading processes: Effects on learning and transfer. *Learning and Instruction, 9,* 109–142.

Couzijn, M., and Rijlaarsdam, G. (2005). Learning to write instructive texts by reader observation and written feedback. *Learning to Write: Reader Observation, 14,* 209–240.

Crawford, L., and Smolkowski, K. (2008). When a "sloppy copy" is good enough: Results of a state writing assessment. *Assessing Writing, 13,* 61–77.

Crawford, L., Helwig, R., and Tindal, G. (2004). Writing performance assessments: How important is extended time? *Journal of Learning Disabilities, 37,* 132–142.

Crawford, L., Tindal, G., and Carpenter II, D. M. (2006). Exploring the validity of the Oregon extended writing assessment. *Journal of Special Education, 40,* 16–27.

Crehan, K. D., and Curfman, M. (2003). Effect on performance of timely feedback on state writing assessments. *Psychological Reports, 92,* 1015–1021.

Davidson, M., Hoekema, P., and Howell, K. W. (2000). Effects of ethnicity and violent content on rubric scores in writing samples. *Journal of Educational Research, 93,* 367–373.

De Ayala, R. J., Dodd, B. G., and Koch, W. R. (1991). Partial credit analysis of writing ability. *Educational and Psychological Measurement, 51,* 103–114.

Deno, S. L., Marston, D., and Mirkin, P. (1982). Valid measurement procedures for continuous evaluation of written expression. *Exceptional Children, 48,* 368–371.

DiStefano, P., and Killion, J. (1984). Assessing writing skills through a process approach. *English Education, 16,* 203–207.

Du, X. (2009). Examining three types of correctional feedback about errors in mechanics and grammar in students with writing difficulties in grades 4–8. (Unpublished doctoral dissertation). University of Minnesota, Minneapolis, MN.

Duke, B. L. (2003). The influence of using cognitive strategy instruction through writing rubrics on high school students' writing self-efficacy, achievement goal orientation, perceptions of classroom goal structures, self-regulation, and writing achievement. (Unpublished doctoral dissertation). University of Oklahoma, Norman, OK.

Engelhard, G., Gordon, B., and Gabrielson, S. (1991). The influences of mode of discourse, experiential demand, and gender on the quality of student writing. *Research in the Teaching of English, 26,* 315–335.

Engelhard, G., Gordon, B., Gabrielson, S., and Walker, E.V. S. (1994). Writing tasks and gender: Influences on writing quality of black and white students. *Journal of Educational Research, 87,* 197–209.

Englebert-Johnson, S. R. (1997). A comparison of English as a foreign language, learning disabled, and regular class pupils on curriculum based measures of reading and written expression and the pupil rating scale revised. (Unpublished doctoral dissertation). University of Minnesota, Minneapolis, MN.

Espin, C., Shin, J., Deno, S. L., Skare, S., Robinson, S., and Benner, B. (2000). Identifying indicators of written expression proficiency for middle school students. *Journal of Special Education, 34,* 140–153.

Espin, C., Wallace, T., Camptell, H., and Lembke, E. S. (2008). Curriculum-based measurement in writing: Predicting the success of high-school students on state-standards tests. *Exceptional Children, 74,* 174–193.

Espin, C. A., De La Paz, S., Scierka, B. J., and Roelofs, L. (2005). The relationship between curriculum-based measures in written expression and quality and completeness of expository writing for middle school students. *Journal of Special Education, 38,* 208–217.

Espin, C. A., Scierka, B. J., Skare, S., and Halverson, N. (1999). Criterion-related validity of curriculum-based measures in writing for secondary school students. *Reading and Writing Quarterly, 14,* 5–27.

Fergusson, C. L., and Fuchs, L. S. (1991). Scoring curriculum-based measurement: A comparison of teachers and microcomputer applications. *Journal of Special Education Technology, 11,* 26–32.

Finlayson, D. S. (1951). The reliability of the marking of essays. *British Journal of Education Psychology, 21,* 126–134.

Fitzgerald, J., and Markham, L. (1987). Teaching children about revision in writing. *Cognition and Instruction, 4,* 3–24.

Franzke, M., Kintsch, E., Caccamise, D., Johnson, N., and Dooley, S. (2005). Summary Street®: Computer support for comprehension and writing. *Journal of Educational Computing Research, 33,* 53–80.

Fuchs, L. S., Fuchs, D., and Hamlett, C. L. (1989). Computers and curriculum-based measurement: Teacher feedback systems. *School Psychology Review, 18,* 112–125.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., and Allinder, R. M. (1991a). Effects of expert system advice within curriculum-based measurement on teacher planning and student achievement in spelling. *School Psychology Review, 20,* 49–66.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., and Allinder, R. M. (1991b). The contribution of skill analysis to curriculum-based measurement in spelling. *Exceptional Children, 57,* 443–453.

Gabrielson, S., Gordon, B., and Engelhard, G. (1995). The effects of task choice on the quality of writing obtained in a statewide assessment. *Applied Measurement in Education, 8,* 273–290.

Gansle, K. A., Noell, G. H., Vanderheyden, A. M., Naquin, G. M., and Slider, N. J. (2002). Moving beyond total words written: The reliability, criterion validity, and time cost of alternative measures for curriculum-based measurement in writing. *School Psychology Review, 31,* 477–497.

Gansle, K. A., Vanderheyden, A. M., Noell, G. H., Resetar, J. L., and Williams, K. L. (2006). The technical adequacy of curriculum-based and rating-based measures of written expression for elementary school students. *School Psychology Review, 35,* 435–450.

Gansle, K. A., Noell, G. H., Vanderheyden, A. M., Slider, N. J., Hoffpauir, L. D., and Whitmarsh, E. L. (2004). An examination of the criterion validity and sensitivity to brief intervention of alternate curriculum-based measures of writing skill. *Psychology in the Schools, 41,* 291–300.

Gearhart, M., Herman, J. L., Baker, E. L., and Whitaker, A. K. (1992). *Writing portfolios at the elementary level: A study of methods for writing assessment* (CSE Tech. Rep. No. 337). Los Angeles, CA: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Gearhart, M., Herman, J. L., Novak, J. R., and Wolf, S. A. (1995). Toward the instructional utility of large-scale writing assessment: Validation of a new narrative rubric. *Assessing Writing, 2,* 207–242.

Godshalk, F. I., Swineford, F., and Coffman, W. E. (1966). *The measurement of writing ability.* New York, NY: College Entrance Examination Board.

Graham, S., and Dwyer, A. (1987). Effects of the learning disability label, quality of writing performance, and examiner's level of expertise on the evaluation of written products. *Journal of Learning Disabilities, 20,* 317–318.

Graham, S., and Leone, P. (1987). Effects of behavioral disability labels, writing performance, and examiner's expertise on the evaluation of written products. *Journal of Experimental Education, 55,* 89–94.

Guastello, E. F. (2001). Parents as partners: Improving children's writing. *Twenty-third Yearbook of the College Reading Association,* 279–293.

Gunel, M., Hand, B., and McDermott, M. A. (2009). Writing for different audiences: Effects on high-school students' conceptual understanding of biology. *Learning and Instruction, 19,* 354–367.

Haladyna, T., and Hess, R. (1999–2000). An evaluation of conjunctive and compensatory standard-setting strategies for test decisions. *Educational Assessment, 6,* 129–153.

Hales, L. W., and Tokar, E. (1975). The effect of the quality of preceding responses on the grades assigned to subsequent responses to an essay question. *Journal of Educational Measurement, 12,* 115–117.

Hebert, M., Graham, S., and Harris, K. (2010). *Predicting writing quality for struggling writers across four genres.* Poster presented at the Embracing Inclusive Approaches for Children and Youth with Special Education Needs Conference, Riga, Latvia.

Hedeker, L. (1997). The effects of month of birth and gender on elementary reading and writing fluency scores using curriculum based measurement. (Unpublished master's thesis). University of Northern British Columbia, Vancouver, Canada.

Herman, J. L., Gearhart, M., and Baker, E. L. (1993). Assessing writing portfolios: Issues in the validity and meaning of scores. *Educational Assessment, 1,* 201–224.

Hogan, T. P., and Mishler, C. (1980). Relationships between essay tests and objective tests of language skills for elementary school students. *Journal of Educational Measurement, 17,* 219–227.

Hollenbeck, K., Tindal, G., and Almond, P. (1999). Reliability and decision consistency: An analysis of writing mode at two times on a statewide test. *Educational Assessment, 6,* 23–40.

Holliway, D., and McCutchen, D. (2004). Audience perspective in young writers' composing and revising. In L. Allal, L. Chanquoy, and P. Largy (Eds.), *Revision: Cognitive and instructional processes* (pp. 87–101). Boston, MA: Kluwer.

Howell, K. W., Bigelow, S. S., Moore, E. L., and Evoy, A. M. (1993). Bias in authentic assessment. *Diagnostique, 19,* 387–400.

Hubbard, D. D. (1996). Technical adequacy of formative monitoring systems: A comparison of three curriculum-based indices of written expression. (Unpublished doctoral dissertation). University of Oregon, Eugene, OR.

Hughes, D. C., Keeling, B., and Tuck, B. F. (1980). The influence of context position and scoring method on essay scoring. *Journal of Educational Measurement, 17,* 131–135.

Hunter, D. M., Jones, R. M., and Randhawa, B. S. (1996). The use of holistic versus analytic scoring for large-scale assessment of writing. *Canadian Journal of Program Evaluation, 11,* 61–85.

Hwang, P. (1930). Errors and improvement in rating English compositions by means of a composition scale. *Contributions to Education, 417,* 1–67.

Jewell, J. (2003). The utility of curriculum-based measurement writing indices for progress monitoring and intervention. (Unpublished doctoral dissertation). Northern Illinois University, Dekalb, IL.

Jewell, J., and Malecki, C. K. (2005). The utility of CBM written language indices: An investigation of product-dependent, product-independent, and accurate-production scores. *School Psychology Review, 34,* 27–44.

Johnson, R. L., Penny, J., Fisher, S., and Kuhs, T. (2003). Score resolution: An investigation of the reliability and validity of resolved scores. *Applied Measurement in Education, 16,* 299–322.

Johnson, R. L., Penny, J., Gordon, B., Shumate, S. R., and Fisher, S. P. (2005). *Language Assessment Quarterly, 2,* 117–146.

Kan, A. (2007). Effects of using a scoring guide on essay scores: Generalizability theory. *Perceptual and Motor Skills, 105,* 891–905.

Kintsch, E., Steinhart, D., Stahl, G., Matthews, C., and Lamb, R. (2000). Developing summarization skills through the use of LSA-based feedback. *Interactive Learning Environments, 8,* 87–109.

Klein, J., and Taub, D. (2005). The effect of variations in handwriting and print on evaluation of student essays. *Assessing Writing, 10,* 134–148.

Koretz, D., Stecher, B., Klein, S., and McCaffrey, D. (1994). The Vermont portfolio assessment program: Findings and implications. *Educational Measurement: Issues and Practice, 13,* 5–16.

Kuhlemeier, H., and van den Bergh, H. (1997). Effects of writing instruction and assessment on functional composition performance. *Assessing Writing, 4,* 203–223.

Lane, K. L., Harris, K. H., Graham, S., Driscoll, S., Sandmel, K., Morphy, P., Hebert, M., and House, E. (In press). Self-regulated strategy development at tier 2 for second-grade students with writing and behavioral difficulties: A randomized controlled trial. *Journal of Research on Educational Effectiveness.*

Lehmann, R. H. (1990). Reliability and generalizability of ratings of compositions. *Studies in Educational Evaluation, 16,* 501–512.

LeMahieu, P., Gitomer, D., and Eresh, J. (1995). Portfolios in large-scale assessment: Difficult but not impossible. *Educational Measurement: Issues and Practice, 14,* 11–28.

Lembke, E., Deno, S. L., and Hall, K. (2003). Identifying and indicator of growth in early writing proficiency for elementary students. *Assessment for Effective Intervention, 28*(3–4), 23–35.

Lumbelli, L., Paoletti, G., and Frausin, T. (1999). Improving the ability to detect comprehension problems: From revising to writing. *Learning and Instruction, 9,* 143–166.

MacArthur, C. A., Schwartz, S. S., and Graham, S. (1991). Effects of a reciprocal peer revision strategy in special education classrooms. *Learning Disabilities Research, 6,* 201–210.

Malecki, C. K., and Jewell, J. (2003). Developmental, gender and practical considerations in scoring curriculum-based measurement writing probes. *Psychology in the Schools, 40*(4), 379–390.

Marshall, J. C., and Powers, J. M. (1969). Writing neatness, composition errors, and essay grades. *Journal of Educational Measurement, 6,* 97–101.

McMaster, K. L., Du, X., and Petursdotter, A. (2009). Technical features of curriculum-based measures for beginning writers. *Journal of Learning Disabilities, 42,* 41–60.

Moon, T., and Hughes, K. (2002). Training and scoring issues involved in large-scale writing assessments. *Educational Measurement: Issues and Practice, 21,* 15–19.

Moss, P. A., Cole, N. S., and Khampalikit, C. (1982). A comparison of procedures to assess written language skills at grades 4, 7, and 10. *Journal of Educational Measurement, 19,* 37–47.

Mott, M. S., Etsler, C., and Drumgold, D. (2003). Applying an analytic writing rubric to children's hypermedia "narratives." *Early Childhood Research and Practice, 5,* 1–17.

Nolet, V., and McLaughlin, M. (1997). Using CBM to explore a consequential basis for the validity of a state-wide performance assessment. *Diagnostique, 22,* 146–163.

Novak, J. R., Herman, J. L., and Gearhart, M. (1996). Establishing validity for performance-based assessments: An illustration for collections of student writing. *Journal of Educational Research, 89,* 220–233.

Olson, V. L. B. (1990). The revising processes of sixth-grade writers with and without peer feedback. *Journal of Educational Research, 84,* 22–29.

Page, E. B., and Paulus, D. H. (1968). *The analysis of essays by computer.* Washington, DC: Office of Education, U.S. Department of Health, Education, and Welfare.

Paquette, K. R. (2009). Integrating the 6+1 writing traits model with cross-age tutoring: An investigation of elementary students' writing development. *Literacy Research and Instruction, 48,* 28–38.

Parker, R. I., Tindal, G., and Hasbrouck, J. (1991a). Countable indices of writing quality: Their suitability for screening-eligibility decisions. *Exceptionality, 2,* 1–17.

Parker, R. I., Tindal, G., and Hasbrouck, J. (1991b). Progress monitoring with objective measures of writing performance for students with mild disabilities. *Exceptional Children, 58,* 61–73.

Penny, J., Johnson, R. L., and Gordon, B. (2000a). The effect of rating augmentation on inter-rater reliability: An empirical study of a holistic rubric. *Assessing Writing, 7,* 143–164.

Penny, J., Johnson, R. L., and Gordon, B. (2000b). Using rating augmentation to expand the scale of an analytic rubric. *Journal of Experimental Education, 68,* 269–287.

Peterson, S., Childs, R., and Kennedy, K. (2004). Written feedback and scoring of sixth-grade girls' and boys' narrative and persuasive writing. *Assessing Writing, 9,* 160–180.

Popp, S. E. O., Ryan, J. M., Thompson, M. S., and Behrens, J. T. (2003). *Operationalizing the rubric: The effect of benchmark selection on the assessed quality of writing.* Paper presented at the annual meeting of the American Educational Research Organization, Chicago, IL.

Prater, D. L., and Bermudez, A. B. (1993). Using peer response groups with limited English proficient writers. *Bilingual Research Journal, 17,* 99–116.

Purves, A. C. (1992). A comparative perspective on the performance of students in written composition. In A. C. Purves (Ed.), *The IEA study of written composition II: Education and performance in fourteen countries* (pp. 129–152). Oxford, UK: Pergamon Press.

Quellmalz, E. S., Capell, F. J., and Chou, C. (1982). Effects of discourse and response mode on the measurement of writing competence. *Journal of Educational Measurement, 19,* 241–258.

Reynolds, C. J., Hill, D. S., Swassing, R. H., and Ward, M. E. (1988). The effects of revision strategy instruction on the writing performance of students with learning disabilities. *Journal of Learning Disabilities, 21,* 540–545.

Rosenthall, B. D. (2006). Improving elementary-age children's writing fluency: A comparison of improvement based on performance feedback frequency. (Unpublished doctoral dissertation). Syracuse University, Syracuse, NY.

Ross, J. A., Rolheiser, C., and Hogboam-Gray, A. (1999). Effects of self-evaluation training on narrative writing. *Assessing Writing, 6,* 107–132.

Russell, M. (1999). Testing on computers: A follow-up study comparing performance on computer and on paper. *Educational Policy Analysis Archives, 7,* 1–47.

Russell, M., and Haney, W. (1997). Testing writing on computers: An experiment comparing student performance on tests conducted via computer and via paper-and-pencil. *Education Policy Analysis Archives, 5,* 1–20.

Russell, M., and Plati, T. (2000). *Mode of administration effects on MCAS composition performance for grades four, eight, and ten.* A report submitted to the Massachusetts Department of Education by the National Board on Educational Testing and Public Policy.

Russell, M., and Tao, W. (2004a). Effects of handwriting and computer-print on composition scores: A follow-up to Powers, Fowles, Farnum, and Ramsey. *Practical Assessment, Research and Evaluation, 9.* Retrieved from http://PAREonline,net/getvn.asp?v=9andn=1

Russell, M., and Tao, W. (2004b). The influence of computer-print on rater scores. *Practical Assessment Research and Evaluation, 9,* 1–17.

Schunk, D. H., and Swartz, C. W. (1993a). Goals and progress feedback: Effects on self-efficacy and writing achievement. *Contemporary Educational Psychology, 18*(3), 337–354.

Schunk, D. H., and Swartz, C. W. (1993b). Writing strategy instruction with gifted students: Effects of goals and feedback on self-efficacy and skills. *Roeper Review, 15,* 225–230.

Sevigny, S., Savard, D., and Beaudoin, I. (2009). Comparability of writing assessment scores across languages: Searching for evidence of valid interpretations. *International Journey of Testing, 9,* 134–150.

Sheppard, E. M. (1929). The effect of quality of penmanship on grades. *Journal of Educational Research, 19,* 102–105.

Shermis, M., Burstein, J., and Bliss, L. (2004). *The impact of automated essay scoring on high stakes writing assessments.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Shohamy, E., Gordon, C., and Kraemer, R. (1992). The effect of rater background and training on the reliability of direct writing tests. *Modern Language Journal, 76,* 28–33.

Soloff, S. (1973). Effect of non-content factors on the grading of essays. *Graduate Research in Education and Related Disciplines, 6,* 44–54.

Steinhardt, D. J. (2001). Summary Street: *An intelligent tutoring system for improving student writing through the use of latent semantic analysis.* (Unpublished dissertation). University of Colorado, Boulder, CO.

Stevens, J. J., and Clauser, P. (1996). *Longitudinal examination of a writing portfolio and the ITBS.* Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY.

Stewart, M. F., and Grobe, G. H. (1979). Syntactic maturity, mechanics of writing, and teachers' quality ratings. *Research in the Teaching of English, 13,* 207–215.

Swartz, C., Hooper, S., Montgomery, J., Wakely, M., Kruif, R., Reed, M., Brown, T., Levine, M., and White, K. (1999). Using generalizability theory to estimate the reliability of writing scores derived from holistic and analytic scoring methods. *Educational and Psychological Measurement, 59,* 492–506.

Swartz, R., and Whitney, D. R. (1985). The relationship between scores of the GED writing skills test and on direct measures of writing. GED *Testing Service Research Studies, 6,* 5–12.

Ternezzi, C. (2009). Curriculum-based measures in writing: A school-based evaluation of predictive validity. (Unpublished dissertation). Western Michigan University, Kalamazoo, MI.

Tezci, E., and Dikici, A. (2006). The effects of digital portfolio assessment process on students' writing and drawing performances. *Turkish Online Journal of Educational Technology, 5*(2), article 7.

Tindal, G., and Parker, R. (1989a). Assessment of written expression for students in compensatory and special education programs. *Journal of Special Education, 23,* 169–183.

Tindal, G., and Parker, R. (1989b). Development of written retell as a curriculum-based measure in secondary programs. *School Psychology Review, 13,* 328–343.

Tindal, G., and Parker, R. (1991). Identifying measures for evaluating written expression. *Learning Disabilities Research and Practice, 6,* 211–218.

Underwood, T., and Murphy, S. (1998). Interrater reliability in a California middle school English/language arts portfolio assessment program. *Assessing Writing, 5,* 201–230.

Veal, L. R., and Hudson, S. A. (1983). Direct and indirect measures for large-scale evaluation of writing. *Research in the Teaching of English, 17,* 290–296.

Veal, L. R., and Tillman, M. (1971). Mode of discourse variation in the evaluation of children's writing. *Research in the Teaching of English, 5,* 37–42.

Vellella, J. A. (1996). The effectiveness of curriculum-based measurement on spelling achievement: A comparison of two procedures. (Unpublished master's thesis). Illinois State University, Normal, IL.

Videen, J., Deno, S., and Marston, D. (1982). Correct word sequences: A valid indicator of proficiency in written expression (Research Report No. 84). University of Minnesota, Institute for Research in Learning Disabilities, Minneapolis, MN.

Wade-Stein, D., and Kintsch, E. (2004). Summary Street: Interactive computer support for writing. *Cognition and Instruction, 22,* 333–362.

Watkinson, J. T., and Lee, S. W. (1992). Curriculum-based measures of written expression for learning-disabled and nondisabled. *Psychology in the Schools, 29,* 184–192.

Weissenburger, J. W., and Espin, C. A. (2005). Curriculum-based measures of writing across grade levels. *Journal of School Psychology, 43,* 153–169.

Wise, W. (1992). The effects of revision instruction on eighth graders' persuasive writing. (Unpublished doctoral dissertation). University of Maryland, College Park, MD.

Wiseman, S. (1949). The marking of English compositions in grammar school selection. *British Journal of Educational Psychology, 19,* 200–209.

Wolf, I. (1986). *Meta-analysis: Quantitative methods for research synthesis.* Beverly Hills, CA: Sage.

Wolfe, E. W., Bolton, S., Feltovich, B., and Bangert, A. W. (1996). A study of word processing experience and its effects on student essay writing. *Journal of Educational Computing Research, 14,* 269–283.

Wolfe, E. W., Bolton, S., Feltovich, B., and Niday, D. M. (1996). The influence of student experience with word processors on the quality of essays written for a direct writing assessment. *Assessing Writing, 3,* 123–147.

Wolfe, E. W., Bolton, S., Feltovich, B., and Welch, C. (1993). A comparison of word-processed and handwritten essays from a standardized writing assessment. *ACT Research Report Series,* 1-30.

Carnegie
CORPORATION OF NEW YORK

ALLIANCE FOR
EXCELLENT EDUCATION

Want to receive the latest information on high school policy?
Add your name to the Alliance's mailing list!

The Alliance for Excellent Education promotes high school transformation to make it possible for every child to graduate prepared for postsecondary education and success in life.

A Washington, DC-based national policy and advocacy organization, the Alliance focuses on America's six million most-at-risk secondary school students—those in the lowest achievement quartile—who are most likely to leave school without a diploma or to graduate unprepared for a productive future.

To add your name to the Alliance mailing list, visit http://www.all4ed.org/whats_at_stake/mailinglist.html or fill out the following form and mail it to the Alliance for Excellent Education at 1201 Connecticut Avenue NW, Suite 901, Washington, DC 20036. You may also fax the form to 202-828-0821. If you have questions, call 202-828-0828.

Name _____

Title _____

Organization _____

Address _____

City/State/Zip _____

Phone _____ Fax _____

Email address _____
(Email address is required to receive *Straight A's.*)

Want to receive the latest information on high school policy?
Add your name to the Alliance's mailing list!

The Alliance for Excellent Education promotes high school transformation to make it possible for every child to graduate prepared for postsecondary education and success in life.

A Washington, DC-based national policy and advocacy organization, the Alliance focuses on America's six million most-at-risk secondary school students—those in the lowest achievement quartile—who are most likely to leave school without a diploma or to graduate unprepared for a productive future.

To add your name to the Alliance mailing list, visit http://www.all4ed.org/whats_at_stake/mailinglist.html or fill out the following form and mail it to the Alliance for Excellent Education at 1201 Connecticut Avenue NW, Suite 901, Washington, DC 20036. You may also fax the form to 202-828-0821. If you have questions, call 202-828-0828.

Name _____

Title _____

Organization _____

Address _____

City/State/Zip _____

Phone _____ Fax _____

Email address _____
(Email address is required to receive *Straight A's.*)

ALLIANCE FOR
EXCELLENT EDUCATION

Straight A's focuses on education news and events in Washington, DC, and around the country. The format makes information on national education issues accessible to everyone from elected officials and policymakers to parents and community leaders. Learn about emerging research, promising practices, and policy decisions that are helping to shape secondary school reform in America.

The Alliance publishes cutting-edge reports such as ***Reading Next*** that combine the best research currently available with well-crafted strategies for turning that research into practice.

Place
Postage
Here

Alliance for Excellent Education
1201 Connecticut Avenue, NW
Suite 901
Washington, DC 20036-2605

ALLIANCE FOR
EXCELLENT EDUCATION

Straight A's focuses on education news and events in Washington, DC, and around the country. The format makes information on national education issues accessible to everyone from elected officials and policymakers to parents and community leaders. Learn about emerging research, promising practices, and policy decisions that are helping to shape secondary school reform in America.

The Alliance publishes cutting-edge reports such as ***Reading Next*** that combine the best research currently available with well-crafted strategies for turning that research into practice.

Place
Postage
Here

Alliance for Excellent Education
1201 Connecticut Avenue, NW
Suite 901
Washington, DC 20036-2605