2015

# FORMATIVE ASSESSMENT AND WRITING: A Meta-Analysis

Steve Graham

Michael Hebert

Karen R. Harris

# FORMATIVE ASSESSMENT AND WRITING

## *A Meta-Analysis*

**ABSTRACT**

To determine whether formative writing assessments that are directly tied to everyday classroom teaching and learning enhance students' writing performance, we conducted a meta-analysis of true and quasi-experiments conducted with students in grades 1 to 8. We found that feedback to students about writing from adults, peers, self, and computers statistically enhanced writing quality, yielding average weighted effect sizes of 0.87, 0.58, 0.62, and 0.38, respectively. We did not find, however, that teachers' monitoring of students' writing progress or implementation of the 6 + 1 Trait Writing model meaningfully enhanced students' writing. The findings from this meta-analysis provide support for the use of formative writing assessments that provide feedback directly to students as part of everyday teaching and learning. We argue that such assessments should be used more frequently by teachers, and that they should play a stronger role in the Next-Generation Assessment Systems being developed by Smarter Balanced and PARCC.

Steve Graham
ARIZONA STATE UNIVERSITY

Michael Hebert
UNIVERSITY OF NEBRASKA–LINCOLN

Karen R. Harris
ARIZONA STATE UNIVERSITY

**DURING** this and the last decade, there have been numerous calls for reforming or improving how writing is taught to children and adolescents. For example, in 2003 the National Commission on Writing (NCoW), established by the College Board, released a report entitled *The Neglected "R": The Need for a Writing Revolution* (NCoW, 2003). The basic thesis of this report was that students in schools in the United States are not receiving the writing instruc-

tion they deserve or need. The report from NCoW called for a comprehensive change in the teaching of writing, urging that writing be squarely placed in the center of efforts to reform educational practices in the United States. The report further recommended that students spend more time writing in and out of school, the use of technology for teaching and assessing writing, professional development for all teachers to improve the teaching of writing, and fair and authentic assessments to evaluate explicitly stated state writing standards.

The Carnegie Corporation of New York also funded three reports (Graham, Harris, & Hebert, 2011; Graham & Hebert, 2010; Graham & Perin, 2007a) during this time period designed to specifically address issues that the authors of the reports believed were roadblocks to making writing a part of school reform efforts. In essence, they reasoned that writing was not more prominent in American schools because policy makers believed that educators did not know how to teach writing effectively, writing had little impact on other important aspects of learning, and assessing writing was of little value. The meta-analyses of true and quasi-experiments presented in these three Carnegie reports provided evidence that this was not the case. The first report, *Writing Next* (Graham & Perin, 2007a), identified a variety of instructional practices that enhanced the quality of students' writing. The effective practices identified in this initial report were expanded and further supported with evidence taken from other types of experiments (i.e., single-subject design; qualitative study of exceptional literacy teachers) in subsequent reviews (e.g., Graham & Perin, 2007b; Rogers & Graham, 2008).

*Writing Next*, along with *Writing to Read* (Graham & Hebert, 2010), provided convincing evidence that writing does enhance learning as well as reading. The meta-analyses in these reports confirmed that learning is enhanced when students write about ideas and information presented in content classes (see also Bangert-Drowns, Hurley, & Wilkinson, 2004), comprehension of text is increased when students write about material read, and teaching writing improves how well students read. The final report, *Informing Writing* (Graham, Harris, & Hebert, 2011), provided evidence that classroom-based writing assessment enhances students' writing performance.

These reports, as well as calls for reforms for teaching writing from other quarters (e.g., ACT, 2005; Langer, 2011), were driven in part by ongoing concerns about how well students write and the quality of writing instruction they receive. Although many children are strong writers and many teachers in the United States provide exemplary writing instruction, this is not the norm. Results from the National Assessment of Educational Progress (National Center for Education Statistics, 2012) showed that the nation made some small progress in improving students' writing, but a majority of students have not mastered the skills necessary for proficient or grade-level-appropriate writing. Furthermore, very little writing or writing instruction takes place in a majority of schools in the United States (e.g., Applebee & Langer, 2006; Gilbert & Graham, 2010: Kiuhara, Graham, & Hawken, 2009).

The most recent of the many subsequent calls for reforming writing instruction is the Common Core State Standards (CCSS) for the English Language Arts. Unlike earlier reform efforts, such as No Child Left Behind, writing is central to the goals of CCSS. The grade-level standards specified in CCSS (2010) provide a road map for the writing skills students need to acquire to be college and career ready. These standards stress that writing is not a generic skill, but it involves mastering the use of writing for multiple purposes, including using writing as a tool to support learning in content

classrooms and comprehending text. The potential impact of CCSS is substantial, as all but a handful of states have agreed to adopt them.

Assessment is one of the basic pillars of CCSS. At present, two consortia, Smarter Balanced (http://www.smarterbalanced.org/higher-education) and Partnership for Assessment of Readiness for College and Careers (PARCC; http://www.parcconline.org), are developing assessments aligned with CCSS. For the most part, the assessments being developed by Smarter Balanced and PARCC involve summative and interim computer assessment tests. The purpose of the summative tests is to assess student progress on CCSS objectives at the end of the school year (they are administered during the last 12 weeks of school). The interim tests can be administered throughout the school year to allow teachers to measure students' progress. Both consortia indicated that these tests will help teachers determine whether their students are on track to meet CCSS objectives, allowing educators to compare student performance across classes, schools, districts, and states.

The consortia also indicated that the summative and interim tests provide teachers with timely assessment information, allowing them to tailor or differentiate instruction to students' needs. Thus, it is expected that results from both of these tests will be used in a formative fashion. While definitions of formative assessment vary, it involves collecting information or evidence about student learning, interpreting it in terms of learners' needs, and using it to alter what happens (Wiliam, 2006). It is anticipated that teachers will use the information provided by these assessments to shape the curriculum as well as student learning.

Even though the assessments from Smarter Balanced and PARCC are still under development, a number of concerns have surfaced, including (1) Do the new tests address past concerns that plagued summative assessment in writing (see Graham, Hebert, & Harris, 2011)? (2) Do interim tests actually increase students' achievement? and (3) Are the assessment systems developed by the two consortia failing to capitalize on the promise that formative assessment holds for teaching and learning (Heritage, 2010)? In terms of the latter issue, formative assessment is viewed by many as a process applied by teachers, students, peers, and even computers that provides feedback for making adjustments in everyday teaching and learning (e.g., Assessment Reform Group, 2002; Formative Assessment for Students and Teachers, State Collaborative on Assessment and Student Standards [FAST CASs], 2008; Heritage, 2010; Stiggins, 2005). While the summative and interim tests developed by the two consortia provide information that can be used for formative purposes (e.g., data that teachers can use to determine students' strengths and weaknesses and adapt or differentiate instruction), these tests are not part of everyday learning and instruction nor do they directly involve students as part of the assessment process (see Stiggins, 2005).

Smarter Balanced and PARCC recognized that formative assessment during learning and instruction is important (e.g., the former is developing a digital library of formative assessment practices for teachers), but it is possible that teachers may reduce their and their students' use of such assessments as part of everyday writing practices, as they may view the summative and interim tests as sufficient to drive positive changes in the learning and teaching of writing. Of course, concerns about this problem are less compelling if everyday formative assessments do not improve how well students write.

The primary purpose of this article is to examine whether formative writing assessments that are directly tied to everyday classroom teaching and learning enhance the quality of students' writing. This includes examining the impact of feedback to students on their writing or their progress in learning specific writing skills or strategies. According to Sadler (1989) and others (e.g., Black & Wiliam, 1988), feedback is the critical element in effective formative assessment, as it provides information that is used by students to improve their writing or learning and by teachers to make changes in their instruction. Accordingly, students use feedback about their writing to close the gap between what they write and the desired goal for a better paper. Such feedback can come from adults (including the teacher and peers), a computer, or through self-assessment, whereas desired goals for writing emanate from multiple sources as well, including professional or personal opinions on what constitutes good writing as well as scoring rubrics and guides that specifically define the attributes of good writing. Likewise, students use instructional feedback about their progress in mastering writing skills and strategies obtained through adult, peer, computer, or self-assessment to improve their learning.

It also includes examining the effectiveness of teachers systematically and frequently monitoring students' writing progress in order to make changes in their teaching with the goal of making it more effective (Sadler, 1989). For our review, this took the form of determining the effectiveness of curriculum-based measurement (CBM; Deno, 1985). With CBM, teachers regularly monitor students' writing progress using test stimuli drawn from the annual writing curriculum to determine the progress of the class as well as individual students to determine the success of their instructional efforts and make adjustments in their teaching accordingly. The goal of this approach, as established by Deno, is to produce accurate and meaningful information in the classroom that indexes students' standing and growth, allowing teachers to determine the effectiveness of their instructional programs and modify them, if needed, to produce better instructional programs. Previous reviews have examined the reliability of common CBM writing measures (Graham, Harris, & Hebert, 2011; McMaster & Espin, 2007). In the current review, we examined whether the application of CBM in writing had a positive impact on students.

We were further interested in determining the effectiveness of the 6 + 1 Trait Writing program developed in conjunction with the Northwest Regional Laboratory (Culham, 2003). This program emphasizes writing instruction in which students and teachers analyze writing using a specific set of traits that include ideas, organization, voice, word choice, sentence fluency, conventions, and presentations. These traits are used to analyze one's own writing and others' writing. They also provide a vocabulary and set of criteria for discussing the qualities of a piece of text with others and for planning one's own writing (Coe, Hanita, Nishioka, & Smiley, 2011). While this is more than a formative assessment procedure per se, it encourages formative assessment of one's own and others' writing as part of the life of the classroom. Thus, we decided to include it as part of this review.

To determine whether these classroom-based formative writing assessments were effective, we conducted a meta-analysis to answer the following questions: (1) Does feedback from adults, peers, computers, and self about writing or learning progress enhance the quality of students' writing? (2) Do adult, peer, self, and computer feedback each improve the quality of students' writing? (3) Does teacher monitoring of students' writing progress (i.e., curriculum-based measurement) result in im-

proved student performance? (4) Does implementation of the 6 + 1 Trait Writing program produce students who are better writers?

Meta-analysis is a statistical tool used to summarize the direction and magnitude of the effects obtained in a set of empirical studies examining the same basic phenomena (Lipsey & Wilson, 2001). The meta-analysis reported in this article drew in part on the work done in *Informing Writing* (Graham, Harris, & Hebert, 2011), and the evidence used to answer each question was derived from true and quasi-experiments. Meta-analysis is well suited to answering the four questions above, as it produces an estimate of the effectiveness of a treatment "under conditions that typify studies in the literature" (Bangert-Drowns et al., 2004, p. 34). Moreover, when enough studies are available and variability in the effects of individual studies is greater than variability due to sampling error alone (which was the case for question 1), meta-analysis allows examining the relationship between study features and outcomes.

The meta-analysis reported here differs from *Informing Writing* (Graham, Harris, & Hebert, 2011) in seven important ways. One, we limited this meta-analysis to studies involving children in grades 1 to 8 (*Informing Writing* spanned grades 1 to 12). This is consistent with the scope and purpose of the *Elementary School Journal*. Two, we examined the combined effects of formative assessment procedures that provided feedback to students (question 1). *Informing Writing* looked at the effects of adult, peer, and self-feedback separately. This decision allowed us to apply meta-regression to examine moderating effects of study characteristics for question 1. As a result, we examined the unique contribution of individual variables (e.g., grade) in accounting for variability in study effects, after variability due to other variables (e.g., study quality; structured vs. unstructured feedback) were first controlled.

Three, we specifically examined the effects of computer feedback on students' writing by including two additional studies in this analysis that were not a true or quasi-experiment. In one of these studies, students acted as their own controls, whereas the other study was an ex post facto causal comparative design (see Method). While these studies were not used to answer question 1, they provided us with four studies examining the effects of computer feedback in question 2. We did not calculate an average weighted effect size for a treatment unless there were at least four studies testing it. This rule of thumb has been applied in other meta-analysis in writing (see Graham & Perin, 2007a; Hillocks, 1986).

Four, we expanded the scope of this meta-analysis to include studies testing the effectiveness of the 6 + 1 Trait Writing program. Five, we adjusted the effects for quasi-experiments included in this review to account for possible data clustering due to hierarchical nesting of data (i.e., researchers assigned classes to treatment or control conditions, but then examined student-level effects). This was not done in *Informing Writing*. Six, the quality of each study was assessed, allowing us to make better judgments about the confidence that could be placed in our conclusions. Seven, the search for appropriate studies was updated as well as expanded to include the electronic database of WorldCat. This involved conducting 96 new searches.

We anticipated that studies examining the effects of feedback about students' writing or their learning progress would produce a positive and statistically significant average weighted effect in improving the quality of writing. Writing quality is based on readers' judgment of the overall merit of a paper, taking into consideration factors such as ideation, organization, vocabulary, sentence structure, tone, and so

forth (Graham & Perin, 2007a). These evaluations are quantified on a numerical scale, representing a single overall judgment (holistic score) or a score for each attribute assessed (analytic score).

We expected that feedback would enhance writing quality when all forms of feedback to students (i.e., adult, peer, self, and computer) were examined collectively (question 1) and separately (question 2). These predictions were based on both theory and previous evidence. From a sociocultural viewpoint, such writing feedback involves a reciprocal activity where teacher, writer, peers, or machine work together to improve students' writing (Heritage, 2010). Even when students self-evaluate their own writing, it represents a collaboration of the writer as creator and evaluator (or reader). There is also considerable evidence that feedback has a positive of effect on learning in areas other than writing (e.g., Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Hattie & Timperley, 2007).

We further anticipated that studies testing the effects of progress monitoring in writing via curriculum-based measurement would have a positive impact on students' performance (question 3). Instructional research in related areas (reading and math) has demonstrated that regularly monitoring students' progress improved the quality of teaching and students' achievement (Fuchs & Fuchs, 1986).

Finally, we predicted that studies testing the 6 + 1 Trait Writing program would produce a positive and statistically significant average weighted effect size (question 4). This prediction was based on the same theoretical and empirical justification provided for questions 1 and 2.

## Method

### Study Inclusion/Exclusion Criteria

Studies had to meet the following six criteria to be included in this meta-analysis. The study (1) was a true experiment (random assignment to conditions) or a quasi-experiment, (2) involved students who were in grades kindergarten to grade 8, (3) contained a treatment group that received a writing assessment intervention, (4) included a measure of writing quality at posttest, (5) was presented in English, and (6) contained the statistics necessary to compute a weighted effect size (or statistics were obtained from the authors).

Studies were excluded if they included students in kindergarten to grade 8, but it was not possible to calculate an effect size just for these students. To illustrate, a study with grade 8 and 9 students was excluded if the data for these students could not be disaggregated by grade. Studies were further excluded if attrition was greater than 20% (e.g., Andrade & Boulay, 2003; Collopy, 2008; Crehan & Curfman, 2003). It is generally agreed that attrition rates of 20% or higher are not acceptable and can change the statistical outcome for a study (Stinner & Tennent, 2012).

We made two exceptions to the inclusion criteria presented above. One, we violated inclusion criterion 2 (true or quasi-experiment) when examining the effects of computer-delivered feedback. There were only two studies that met this criterion. As a result, we included computer feedback studies where students served as their own controls as well as investigations that involved ex post facto causal comparative design. In the former, the same students' performance with and without computer feedback was compared. In the latter, students who did and did not receive the

treatment (i.e., computer-delivered feedback) were compared after the fact to determine whether a possible causal relationship exists between the treatment and changes in the quality of students' writing (Griffin, 2000). Such studies are less preferable than true and quasi-experiments, but their inclusion allowed us to compute an average weighted effect for computer feedback, providing an initial exploratory test of the impact of this treatment.

Two, we violated inclusion criterion 4 (measure of writing quality at posttest) for studies that examined the effectiveness of CBM. This form of progress monitoring typically involves more discrete measures of students' writing.

### Search Strategies Used to Locate Studies

To identify possible studies for this meta-analysis, electronic searches were run in multiple databases, including ERIC, PsychINFO, ProQuest, Education Abstracts, WorldCat, and Dissertation Abstracts. The following search terms were paired with writing: assessment, evaluation, portfolio, performance assessment (students perform a task as part of assessment), curriculum-based assessment, curriculum-based measurement, automated essay scoring, computer scoring, analytic quality, holistic quality, word processing, self-assessment, feedback, peer feedback, and 6 + 1 Trait Writing. Close to 7,500 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.

The use of these search terms resulted in a broad search to identify pertinent studies. Terms such as assessment and evaluation when paired with writing yielded a variety of different studies on writing assessment (including studies on formative assessment). Likewise, searching for items using the terms for the two most common measures (holistic and analytic quality) used to assess the primary outcome of interest in this review (i.e., writing quality) increased the likelihood of locating relevant studies. In addition, we conducted more localized and strategic reviews by pairing the term writing with portfolio, performance assessment, curriculum-based assessment, curriculum-based measurement, automated essay scoring, computer scoring, self-assessment, feedback, peer feedback, and 6 + 1 Trait Writing. Lastly, we included word processing as a search term because formative assessments can involve digital forms of writing.

Hand searches were also conducted with the following peer-reviewed journals: *Assessing Writing*, *Journal of Writing Assessment*, *Research in the Teaching of English*, and *Written Communication*. Moreover, previous reviews (Graham, Bollinger, et al., 2012; Graham & Perin, 2007b; Graham, McKeown, Kiuhara, & Harris, 2012; Hillocks, 1986) were examined to identify additional studies. Once a document was obtained, the reference list was searched to identify additional promising studies. Of 539 documents collected, we found 34 papers that contained 35 experiments that met all of the inclusion criteria. These 35 experiments yielded 39 effect sizes.

The most common reason for why an obtained document was not included were (in the following order) the study did not involve a true or quasi-experiment (rejected studies included studies without a control group, descriptive studies, validity and reliability studies, and qualitative studies); the study did not include writing quality as an outcome measure; the document was not a study (instead it was a

discussion piece or a review of literature); and the statistics for calculating an ES were unobtainable.

## Categorizing Studies into Treatment Conditions

**Step 1.** First, each obtained study was read by the first author and placed into one of the following two categories: (1) it met the inclusion criteria, or (2) it did not meet the inclusion criteria. Studies placed into category 2 were read a second time to ensure that they should be excluded. Only one study was reassigned to category 1.

**Step 2.** Studies placed into category 1 were reread to ensure that they met inclusion criteria (all did). At the same time, the first author developed an initial set of subcategories for these investigations (e.g., self-assessment, peer feedback, teacher feedback). This process of reading studies and sorting them into categories was repeated several times, resulting in the following subcategories of studies: impact of feedback (peer or adult), self-assessment, curriculum-based measurement, computer marking systems, and 6 + 1 Trait Writing programs. Once these subcategories were created, all studies, including the ones that were initially excluded (i.e., category 2), were reexamined to determine whether they belonged in their assigned subcategory and whether other subcategories needed to be created. All studies fit their assigned subcategory and no new categories were created.

Reliability of this categorization process was established by having a graduate student in educational psychology read and categorize all studies. There was only one disagreement with the first author, which was resolved through discussion, and the study was categorized as it was originally coded.

## Coding of Study Features

Each study was coded for study characteristics, quality indicators, and statistics needed to calculate effect sizes. Study characteristics included grade, participant type (e.g., struggling writers, English Language Learners, etc.), number of participants, genre of the posttest measure (e.g., narrative, expository, persuasive), brief description of treatment and control conditions, and publication type.

Each study was examined to determine whether nine quality indicators were met. Each indicator was scored as 1 (met) or 0 (not met). Quality indicators included (1) design (true experiment was assigned a score of 1, whereas quasi-experiment, subjects as own control, and nonexperimental comparative design were scored as 0); (2) the control treatment was defined; (3) treatment fidelity established through direct observation; (4) teacher effects controlled (e.g., random assignment of teachers); (5) multiple teachers carried out each condition; (6) total attrition was less than 10% of total sample; (7) total attrition was less than 10%; (8) equal attrition across conditions (i.e., conditions did not differ by more than 5%); (9) pretest equivalence established in quasi-experiments (i.e., conditions did not differ by more than 1 *SD* for the condition with the least variance; this was scored as 0 if there was no pretest); (10) pretest ceiling/floor effects were not evident in quasi-experiments (more than 1 *SD* from floor and ceiling; this was scored as 0 if there was no pretest); and (11) posttest ceiling/floor effects were not evident (more than 1 *SD* from floor and ceiling). A total score was calculated for each study (9 possible points for true experiments, and 11

possible points for quasi-experiments). This was converted to a percentage by dividing the obtained score by total possible points and multiplying by 100%.

Coding for study descriptors and quality indicators were done by the second author. Reliability was established by a graduate student in educational psychology on 50% of the studies. Interrater agreement was 94.3% for all variables.

## Calculation of Effect Sizes

**Basic procedures.** An effect size (ES) was calculated by subtracting the mean score of the control group at posttest ($\bar{X}_T$) from the mean score of the treatment group at posttest ($\bar{X}_C$) and dividing by the pooled standard deviation of the two groups ($s_p$) using the following formula (Lipsey & Wilson, 2001):

$$ES_{sm} = \frac{\bar{X}_T - \bar{X}_C}{s_p}, \; s_p = \sqrt{\frac{(s^2_1)(n_1 - 1) + (s^2_2)(n_2 - 1)}{n_1 + n_2 - 2}}.$$

If a comparable pretest measure was available, the same formula was used, except pretest differences between treatment and control conditions were first adjusted by subtracting mean pretest score for each group from their mean posttest score. All effects were adjusted for small-sample-size bias ($d_{\text{adj}} = d^{\star}\gamma$; $\gamma = 1 - 3/4(n_{tx} + n_{ctrl}) - 9$; Hedges, 1982).

If the statistics needed to compute an ES were missing from a paper, we estimated them from the statistics provided whenever possible. For example, missing standard deviations for covariate or complex factorial designs were estimated by calculating and restoring the variance explained by covariates and other "off-factors" to the study's error term and recalculating the root-mean-squared error (RMSE), or pooled standard deviation, from the composite variance.

As noted earlier, effect sizes were calculated for writing quality in all studies except those involving curriculum-based measurement (where spelling performance was the outcome in most studies). If a study used a holistic measure to assess writing quality (i.e., raters assigned a single score for overall quality), an ES was computed for this measure. If both a holistic and analytic measure (raters assigned separate scores for specific aspects of writing, such as content, organization, vocabulary, mechanics, and so forth) were available, an ES was only computed for the holistic measure. If only an analytic measure was available, a separate ES was computed for each aspect of writing assessed and averaged to produce a single ES (similar to a holistic rating). Finally, if only a norm-referenced outcome measures was available and the score from it was based on the quality or schematic structure of a sample of student's writing, an ES for writing quality was computed.

*Calculating effect sizes for separate subgroups.* As a prelude to calculating some effect sizes, it was necessary to average the performance of two or more groups in each condition (e.g., statistics were reported separately by grade, gender, or type of student). To aggregate such data, a procedure recommended by Nouri and Greenberg (Cortina & Nouri, 2000) was applied. This method estimates an aggregate group or grand mean and provides a correct calculation of the variance by combining the variance within and between groups. First, we calculated the aggregate treatment or control mean as an *n*-weighted average of subgroup means:

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

Next, 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:

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

*Adjusting effect size estimates for clustering within treatments.* The quasi-experiments in this meta-analysis assigned whole classes to treatment or control conditions, and then examined student-level effects. It was necessary to adjust standard errors (*SE*) for these studies, as a portion of the total variance in such studies was likely due to grouping or clustering within treatments, with the total variance representing a sum of group and student variances. We estimated $\delta_T$ by adjusting the conventional effect sizes using the intraclass correlation estimator "ES = $d_T$" recommended by Hedges (2007):

$$d_T = \left(\frac{Y^T_{..} - Y^C_{..}}{S_T}\right)\sqrt{1 - \frac{2(n-1)\rho}{N-2}},$$

where $Y^T_{..}$ is the grand mean of the treatment group, $Y^C_{..}$ is the grand mean for the control group, $S_T$ is the total pooled within-treatment variance, *n* is the number of students within cluster, *N* is the number of students total, and $\rho$ is the intraclass correlation.

The variance of the effect sizes further had to be adjusted to include the variance associated with clustering. The equation for calculating the variance of $d_T$ is normally distributed, and we calculated it using the following equation provided by Hedges (2007):

$$\nu_T = \left(\frac{N^T + N^C}{N^T N^C}\right)(1 + (n-1)\rho) + d^2_T\left(\frac{(N-2)(1-\rho)^2 + n(N-2n)\rho^2 + 2(N-2n)\rho(1-\rho)}{2(N-2)[(N-2) - (N-1)\rho]}\right),$$

where $N^T$ is the total number of students in the treatment group, and $N^C$ is the total number of student in the control group, with additional symbols defined in the previous paragraph.

To use these formulas, it was necessary to impute the intraclass correlations (ICCs), or $\rho$, because they were not reported in any of the obtained studies. As was done by Graham, McKeown, et al. (2012) and Morphy and Graham (2012), ICCs were imputed using ICC estimates for reading comprehension from national studies (Hedges & Hedberg, 2007) that were adjusted to writing quality ICCs, using data from a large multistate study of writing involving a single grade level (Rock, 2007). While we would have preferred using ICCs based on writing data at each grade, such statistics were not available. ICCs based on reading provide a reasonably good match to writing, as students' performance on these two skills are strongly related (Fitzger-

ald & Shanahan, 2000), and we were able to adjust ICCs from Hedges and Hedberg (2007) using the Rock (2007) data.

### Statistical Analysis of Effect Sizes

**Average weighted effect size.** An average weighted effect size was computed for a specific writing assessment treatment when there were at least four or more independent comparisons assessing its effectiveness. Although Graham, McKeown, et al. (2012), Graham and Perin (2007a), and Hillocks (1986) applied the same criteria, it must be recognized that small sample sizes are less reliable than larger ones.

Our meta-analysis employed a weighted random-effects model (weighted to take into account sample size by multiplying each ES by its inverse variance). For each writing assessment treatment, we calculated the mean and confidence interval for the average weighted effect size.

We further calculated two measures of homogeneity ($Q$ and $I^2$) for each average weighted ES. The homogeneity measures allowed us to determine whether variability in the effect sizes for a specific spelling treatment was larger than expected based on sampling error alone. The $Q$ statistic is typically used to determine whether excess variability in effect sizes exists, but it is underpowered when there are relatively few effect sizes (which was the case for the analyses in this review). As a result, we also computed $I^2$, which indicates the percent of variance due to between-subject factors.

When variability in effect sizes was larger than expected based on sampling error alone and there were at least 12 effect sizes computed for the treatment, we examined whether this excess variability could be accounted for by identifiable differences between studies' meta-regression (Lipsey & Wilson, 2001). More specifically, we examined whether excess variability in effects for feedback was related to study quality, grade of students (elementary students in grades 1 to 5 vs. middle-school students in grades 6 to 8), and type of feedback (structured feedback from a rubric, strategy or list vs. unstructured feedback). For structured feedback, input to students was directed in advance to particular aspects of students' writing or performance, whereas input from unstructured feedback was not predetermined or directed. We anticipated that weaker studies and structured feedback would yield larger effects. We also anticipated that feedback would be more effective with older students, as they would be better able to take advantage of such information due to their greater skill and experience with writing.

## Results

Table 1 contains information on the studies testing the effectiveness of each writing assessment treatment. Treatment categories are arranged from those assessing the effectiveness of various forms of feedback (i.e., feedback from adults, peers, self, and machine) to those evaluating curriculum-based measurement and 6 + 1 Trait Writing programs. Studies included under each writing assessment treatment report the following information: reference, publication type, grade, participant type, number of participants in the study, brief description of treatment and control condition, genre(s) of writing emphasized, study quality score (percentage of quality indicators met by a study), and ES. Table 2 includes the number of studies, average weighted ES,

Table 1. Study Descriptors Listed by Comparison

| | Experiment | Pub Type | Grade | Student Type | *n* | Genre | Quality Score | ES |
|---|---|---|---|---|---|---|---|---|
| Effects of adult feedback on students' writing quality: | | | | | | | | |
| Rosenthal (2006) | Q | D | 3 | A & AA | 45 | E | .64 | .23 |
| Guastello (2001) | Q | J | 4 | F | 167 | N | .64 | 1.01 |
| Schunk & Swartz (1993a, Study 2) | T | J | 4 | F | 20 | V | .82 | .86 |
| Schunk & Schwartz (1993b) | T | J | 4 | G | 22 | V | .73 | .92 |
| Schunk & Swartz (1993a, Study 1) | T | J | 5 | F | 30 | V | .91 | .67 |
| Wolter (1975) | T | D | 6 | F | 27 | S | .55 | .90 |
| Lumbelli et al. (1999) | T | J | 6 | NS | 28 | I | .64 | .83 |
| Effects of peer feedback on students' writing quality: | | | | | | | | |
| Prater & Bermudez (1993) | T | J | 4 | ELL | 46 | N | .73 | .15 |
| Philippakos (2012) | T | D | 4–5 | F | 97 | P | .82 | .31 |
| MacArthur et al. (1991) | Q | J | 4–6 | LD | 29 | N | .73 | 1.33 |
| Boscolo & Ascorti (2004) | Q | J | 4, 6, 8 | F | 122 | N | .78 | .97 |
| Holliway (2004) | T | J | 5 | F | 55 | E | .50 | .58 |
| Olson (1990) | Q | J | 6 | F | 42 | N | .64 | .71 |
| Benson (1979) | Q | D | 6–8 | F | 288 | S & P | .64 | .36 |
| Wise (1992) | Q | D | 8 | F | 88 | P | .33 | .63 |
| Effects of self-assessment on students' writing quality: | | | | | | | | |
| Paquette (2009) | Q | J | 2 | F | 85 | E | .55 | .70 |
| Andrade et al. (2008) | Q | J | 3 & 4 | F | 116 | S & P | .67 | .85 |
| Young (2000) | Q | D | 4 | F | 161 | NS | .82 | .82 |
| Guastello (2001) | Q | J | 4 | F | 167 | N | .64 | 1.22 |
| Ross et al. (1999) | Q | J | 4–6 | F | 296 | N | .73 | .17 |
| Olson (1990) | Q | J | 6 | F | 42 | N | .54 | .18 |
| Fitzgerald & Markham (1987) | T | J | 6 | F | 30 | S | .82 | .31 |
| Wolter (1975) | T | D | 6 | F | 27 | S | .55 | 1.25 |
| Reynolds et al. (1988) | Q | J | 6–8 | LD | 53 | S | .45 | .15 |
| Wise (1992) | Q | D | 8 | F | 87 | P | .33 | .62 |
| Effects of adult, peer, and self-assessment on students' writing quality: | | | | | | | | |
| Meyer et al. (2010) | Q | J | 4–6 | F | 296 | N | .55 | .29 |
| Effects of computer feedback on students' writing quality | | | | | | | | |
| Wade-Stein & Kintsch (2004) | Q | J | 6 | F | 52 | SUM | .22 | .42 |
| Caccamise et al. (2007) | Q | J | 6–9 | F | 140 | SUM | 0 | .38 |
| Holman (2011) | Q | D | 8 | F | 160 | NS | .43 | .44 |
| Franzke et al. (2005) | T | J | 8 | F | 111 | SUM | .711 | .31 |
| Effects of progress monitoring (curriculum-based measurement) on students' writing: | | | | | | | | |
| Vellella (1996) | Q | D | 2 | F | 91 | SP | .67 | .18 |
| Fuchs et al. (1991a) | T | J | 2–8 | LD | 60 | SP | 1 | .28 |
| Jewell (2003) | T | D | 3, 5, 8 | F | 257 | S | .89 | .12 |
| Fuchs et al. (1991b) | T | J | 3–9 | LD | 100 | SP | .91 | .26 |
| Fuchs et al. (1989) | T | J | ELEM | LD | 54 | SP | 1 | .26 |

Table 1. (*Continued*)

| | Experiment | Pub Type | Grade | Student Type | *n* | Genre | Quality Score | ES |
|---|---|---|---|---|---|---|---|---|
| Effects of 6 + 1 Trait Writing model on students' writing quality: | | | | | | | | |
| Adler (1998) | Q | D | 3 | F | 81 | NS | .36 | .18 |
| Kozlow & Bellamy (2004) | T | TR | 3–6 | F | 1,592 | N, E, P | .82 | .10 |
| Arter (1994) | Q | CP | 5 | F | 132 | NS | .45 | .19 |
| Coe et al. (2011) | T | TR | 5 | F | 4,161 | NS | .73 | .04 |

Note.—Q = quasi-experiment, T = true experiment, ES = effect size, J = journal, D = dissertation, TR = technical report, CP = conference paper, F = full range, G = gifted, A = average, AA = above average, LD = learning disabled, ELEM = unspecified elementary grades, NS = not specified, V = varied, S = story, E = essays, N = narrative, I = informative/descriptive, P = persuasive, SUM = summary, SP = spelling.

confidence interval, standard error, and statistical significance for each treatment as well the two heterogeneity measures ($Q$ and $I^2$).

## Quality of Research

As can be seen in Table 1, the quality of studies varied widely, with some studies meeting all of the quality indicators and one study meeting none of them. In terms of quality of studies by category of treatment, progress monitoring met the highest percentage of quality indicators (91%), followed by adult feedback (70%), peer feedback (65%), self-assessment (62%), 6 + 1 Trait Writing model (59%), and computer feedback (30%). With a few exceptions, most studies did not evidence problems with reliability of measures, pretest equivalence, or ceiling/floor effects at pretest and posttest. With the exception of studies investigating progress monitoring, researchers rarely provided evidence confirming that the independent variable or treatment was implemented as intended. Attrition and providing an adequate description of the control condition was a problem in studies testing computer feedback as well as the 6+1 Trait Writing model. Researchers did not adequately control for teacher

Table 2. Average Weighted Effect Sizes and Confidence Intervals for Writing Assessment Treatments

| | | | | Test of Null Hypothesis | | Heterogeneity | |
|---|---|---|---|---|---|---|---|
| Writing Intervention | Studies | Effect Size | Confidence Interval | Variance | *p*-Value | *Q*-Value | $I^2$ |
| All studies involving feedback | 27 | .61 | (.42, .79) | .01 | <.001 | 106.39** | 77.56 |
| Adult feedback | 7 | .87 | (.62, 1.11) | .02 | <.001 | 3.39 | .00 |
| Peer feedback | 8 | .58 | (.35, .82) | .01 | <.001 | 13.49 | 48.10 |
| Self-assessment | 10 | .62 | (.34, .90) | .02 | <.001 | 36.49** | 75.34 |
| Computer feedback | 4 | .38 | (.17, .59) | .01 | .001 | .22 | .00 |
| Progress monitoring | 5 | .18 | (-.01, .36) | .01 | .06 | .56 | .00 |
| 6 + 1 Trait Writing model | 4 | .05 | (-.01, .11) | .001 | .08 | .72 | .00 |

**$p < .001$.

effects in a third or more of the studies testing adult, peer, and self-feedback. While true experiments were commonly used to test adult feedback and progress monitoring, this was not the case for the other treatments.

## Question 1: Does Feedback about Students' Writing or Their Learning Progress Enhance the Quality of Students' Writing?

We calculated 27 effect sizes for writing quality from 25 papers where the effectiveness of feedback in writing was tested. In order to avoid inflating sample size and violating the assumption of independence of data (Wolf, 1986), it is generally recommended that only one effect size for each study is calculated when computing an average weighted ES or conducting a meta-regression. One paper (Schunk & Swartz, 1993a) included two separate investigations, so an ES was calculated for each study. We also calculated two effect sizes for Olson (1990), as her study included multiple treatment conditions with different control conditions.

Seven of the studies included in this analysis assessed the effectiveness of adult feedback (Guastello, 2001; Lumbelli, Paoletti, & Frausin, 1999; Rosenthal, 2006; Schunk & Swartz, 1993a [studies 1 and 2]; Schunk & Swartz, 1993b; Wolter, 1975). Three of the studies involved teachers providing feedback to students on their progress in learning to write paragraphs. One study assessed the impact of teacher feedback to students on their writing, with one investigation providing students with written teacher feedback on correct word sequence, spelling, and total words written. In another study, the experimenter provided students with feedback using a specific scoring form to structure the feedback provided. In another study, parents gave feedback to their child on their writing; they received training on how to provide such feedback. In the final study, students revised text while receiving verbal recorded feedback from an adult on how to do so.

Eight studies included in this analysis assessed the impact of peer feedback. In six of these investigations, peers both gave feedback to one or more peers on their writing and received feedback about their own writing from one or more classmates (Benson, 1979; Boscolo & Ascorti, 2004; MacArthur, Schwartz, & Graham, 1991; Olson, 1990; Prater & Bermudez, 1993; Wise, 1992). In two studies (Holliway, 2004; Philippakos, 2012), students gave feedback to their peers on their writing, but did not receive such feedback themselves. The methods for providing feedback in the eight studies were varied and included (1) a direction to meet with another classmate and provide feedback on their writing, (2) specifying specific aspects of writing that students were to focus on when providing peer feedback (e.g., unclear parts, gaps in content, adequacy of description), (3) teaching students to use a rubric or scale for providing feedback, and (4) teaching selected strategies for providing feedback (these typically focused on noting positive aspects of the classmate's writing and providing feedback on particular attributes such as clarity or completeness).

Ten comparisons tested the impact of self-assessment on students' writing (Andrade, Du, & Wang, 2008; Fitzgerald & Markham, 1987; Guastello, 2001; Olson, 1990; Paquette, 2009; Reynolds, Hill, Swassing, & Ward, 1988; Ross, Rolheiser, & Hogaboam-Gray, 1999; Wise, 1992; Wolter, 1975; Young, 2000). In all 10 studies, students received either minimal or more intensive instruction in how to self-assess and revise their writing. This most often included instruction in how to use a rubric to score their writing or scoring form ($N = 7$), but it also included teaching them how

to carry out specific revising tactics such as substituting, adding, deleting, or moving text to improve their writing.

It should be noted that we placed the study conducted by Paquette (2009) in the self-assessment category. In this investigation, fourth-grade tutors taught second-grade students how to use a rubric to assess writing produced by the tutor. In our estimation, this was not a peer feedback study, as the purpose of these assessments was not to improve the tutors' papers, but rather to strengthen the tutee's writing self-assessment skills. The process of teaching a younger child how to conduct such an assessment should also improve the tutors' self-assessment skills. Thus, in contrast to Graham, Harris, and Hebert (2011), we did not categorize the Paquette study as a peer feedback investigation, and the ES we computed was based on changes in the tutors' and tutees' writing performance.

Of the five remaining studies that were included in this analysis, one investigation (Meyer, Abrami, Wade, Aslan, & Deault, 2010) assessed the effects of a combination of teacher, peer, and self-feedback of writing provided as part of an electronic portfolio system. The other four studies tested the effects of computer feedback on students' writing. Three of these studies tested the impact of Summary Street, a computer program that provided students with feedback on summaries they wrote (Caccamise, Franzke, Eckhoff, Kintsch, & Kintsch, 2007; Franzke, Kintsch, Caccamise, Johnson, & Dooley, 2005; Wade-Stein & Kintsch, 2004). The fourth study (Holman, 2011) tested the effectiveness of *MY Access!* This was a Web program from Vantage Learning that provided students with feedback on their writing.

Our meta-analysis of the 27 feedback comparisons yielded a statistically significant average weighted ES of 0.61. All 27 comparisons resulted in a positive ES. Variability in effect sizes was statistically greater than sampling error alone (see *Q* in Table 2), and the $I^2$ statistic indicated that 78% of the variance was due to between-study factors (see Table 2).

A funnel plot of precision was plotted to examine possible publication bias. There did not appear to be publication bias when examining this plot with observed and imputed effect sizes using Duval and Tweedie's Trim and Fill procedure. In addition, the Begg and Mazumdar Rank Correlation Test was not statistically significant ($p > .07$), and the classic fail-safe test showed that 1,112 missing studies would need to be collected to nullify the statistically significant effect.

Given that variability in the 27 effect sizes was greater than sampling error alone, we conducted a preplanned meta-regression to determine whether quality of studies, grade level (grades 1 to 5 vs. grades 6 to 8), and feedback structure (i.e., structured vs. unstructured) accounted for some of this excess variability. The meta-regression involved a mixed-effects model with maximum likelihood estimates using macros developed for SPSS. We assumed that in addition to a random effect due to sampling error, there was a systematic component to the variance between studies that could be explained by between-studies factors. The macros added a random effect variance component and recalculated the inverse variance weight before refitting the model (Lipsey & Wilson, 2001). The three predictor variables were entered as a single block. The average weighted effect size for the 27 studies in the meta-regression was 0.63. The analysis (see Table 3) did not explain excess variability in effect sizes ($Q$-value = 1.37, $df[Q] = 3$, $p = .712$). The constant was statistically significant, however, indicating an average ES of 0.81 across grade levels after accounting for variability due to

Table 3. Meta-Regression of Effect Size on Specified Study Characteristics

| | Mean ES | *R*-Square | *k* | | | |
|---|---|---|---|---|---|---|
| Descriptives | .63 | .06 | 26 | | | |
| | *Q* | *df* | *p* | | | |
| Homogeneity analysis: | | | | | | |
| Model | 1.37 | 3 | .712 | | | |
| Residual | 23.00 | 22 | .402 | | | |
| Total | 24.37 | 25 | .498 | | | |
| | | | 95% CI | | | |
| | B | *SE* | Lower | Upper | *Z* | *p* |
| Regression coefficients: | | | | | | |
| Constant | .81 | .20 | .44 | 1.20 | 4.02 | <.001 |
| Structured | −.14 | .21 | −.56 | .27 | −.68 | .500 |
| ELEM vs. MS | −.19 | .20 | −.59 | .21 | −.94 | .350 |
| Quality | −.03 | .55 | −1.11 | 1.05 | −.06 | .956 |

Note.—ELEM = elementary grades; MS = middle school; structured = dummy code comparing studies incorporating unstructured feedback (0) to studies that incorporated structured feedback (1); ELEM vs. SEC = dummy code comparing studies done with elementary grade students (0) to studies with middle-school students (1); quality = proportion of study quality variables met by the study.

study quality and structured feedback. None of the individual variables made unique or statistically significant contributions to the model.

## Question 2: Do Adult, Peer, Self, and Computer Feedback Each Improve the Quality of Students' Writing?

**Adult feedback.** As noted earlier, adult feedback involved the teacher as well as other adults giving students feedback on their writing as well as teachers providing students with feedback on their progress in learning a writing strategy. Collectively, the seven studies testing the effectiveness of adult feedback yielded a statistically significant average weighted ES of 0.87. Variability in effect sizes was not statistically greater than sampling error alone (see *Q* in Table 2), and the $I^2$ statistic indicated that none of the variance was due to between-study factors (see Table 2).

**Peer feedback.** Studies assessing the effectiveness of peer feedback included two types of studies: (1) peers gave and received feedback about their writing from other classmates ($N = 6$) or (2) peers gave feedback to their peers about their writing ($N = 2$). Together all eight studies produced a statistically significant average weighted ES of 0.58. Variability in effect sizes was not statistically greater than sampling error alone (see *Q* in Table 2), and the $I^2$ statistic indicated that 48% of the variance was due to between-study factors (see Table 2).

**Self-assessment.** Students were taught to self-assess their own writing in 10 investigations. These studies yielded a statistically significant average weighted ES of 0.62. Variability in effect sizes was statistically greater than sampling error alone (see *Q* in Table 2), and the $I^2$ statistic indicated that 75% of the variance was due to between-study factors (see Table 2) .

**Computer feedback.** The four studies testing the effects of computer feedback yielded a statistically significant average weighted ES of 0.38. Variability in effect sizes was not statistically greater than sampling error alone (see *Q* in Table 2), and the $I^2$

statistic indicated that none of the variance was due to between-study factors (see Table 2).

### Question 3: Does Teacher Monitoring of Students' Writing Progress Result in Improved Student Performance?

We located five studies where teachers monitored students' progress on one or more writing variables. The outcome of interest in these studies was not the CBM assessments. Instead effect sizes were computed with other broader outcome measures administered at posttest (see below). Four of these studies (Fuchs, Fuchs, & Hamlett, 1989; Fuchs, Fuchs, Hamlett, & Allinder, 1991a, 1991b; Vellella, 1996) involved teachers tracking students' spelling progress weekly over a 3- to 4-month period. The outcome measure in these four studies was performance on a norm-referenced spelling test. In the fifth study (Jewell, 2003), teachers monitored weekly changes in students' performance on a variety of measures (e.g., words written, spelling, correct word sequence) over a 3-month period. The outcome measure in this study was the quality of students' writing on the state writing test.

While all five studies produced a positive effect, collectively they did not produce a statistically significant average weighted ES (see Table 2). The effect was small (0.18) and would not be considered substantially important using the criteria established by the What Works Clearinghouse (see Graham, Bollinger, et al., 2012). This small effect did not appear to be a result of poor implementation, as fidelity of implementation was strong across all five studies.

### Question 4: Does Implementation of the 6 + 1 Trait Writing Model Produce Students Who Are Better Writers?

Four studies were located that examined the effectiveness of the 6 + 1 Trait Writing program. Three of these studies (Arter, 1994; Coe et al., 2011; Kozlow & Bellamy, 2004) were conducted by researchers at the Northwest Regional Educational Laboratory (NWREL), the developers of this program. In these three studies, teachers received considerable professional development training from the NWREL on how to implement the program, classes or schools were randomly assigned to treatment and control, the treatment lasted most of a school year, and with one exception (Arter) the studies were relatively large (involving between 76 to close to 200 teachers). The investigation by Arter included six teachers randomly assigned to either treatment or control conditions. Two of these studies involved fifth-grade students (Arter; Coe et al.), whereas the other study (Kozlow & Bellamy) included third to sixth graders.

The fourth study testing the effectiveness of the 6 + 1 Trait Writing program was a doctoral dissertation conducted by Adler (1998) with third-grade children. Two teachers received training in how to implement the program and applied the program over a 4-month period. Their gains in writing quality were compared to the gains made by students in two classes that did not receive professional development in the program.

In three of the studies (Adler, 1998; Arter, 1994; Kozlow & Bellamy, 2004), the program appeared to be implemented with generally good fidelity. In the largest and best-designed study, the authors indicated that the level of implementation was unclear. It should further be noted that we did not include a study by Paquette (2009) in

the analysis presented below, as it assessed the effectiveness of a cross-age tutoring program conducted with fourth- and second-grade children. While it focused on the 6 + 1 Trait Writing, it was not a test of the basic teacher implemented model, as was the case with the other four studies.

All four studies had a positive effect, but collectively they yielded an average weighted ES of only 0.05, which was not statistically significant. Variability in effect sizes was not statistically greater than sampling error alone (see *Q* in Table 2), and the $I^2$ statistic indicated that none of the variance was due to between-study factors (see Table 2).

We reran this analysis by winsorizing the sample size when computing the ES for Coe et al. (2011) so that it did not exert an undue influence on the analysis. Typically, we would limit the sample size for the control and experimental condition of Coe et al. by following Tukey's (1977) recommendation of confining an extreme observation to three times the interquartile range above the 75th percentile of the distribution of all related observations. However, we were not able to calculate the 75th percentile with just three data points (the sample sizes from the remaining three studies). Consequently, we decided to winsorize the sample size for Coe et al. by using the sample size from the next largest study (i.e., Kozlow & Bellamy, 2004). When we reran the analysis, the average weighted ES increased to 0.08 (confidence interval = −0.03 to .18), but this effect was still not statistically significant ($p = .17$).

In computing an ES for the Coe et al. (2011) investigation, we had to convert standard errors at posttest to standard deviations. We obtained a different effect than the one reported by Coe et al. (0.041 vs. 0.109). This may have been the result of differences in how effects were calculated. First, our calculation of standard deviations may have not taken into account all of the factors involved in computing standard errors in Coe et al. Second, when calculating an ES, Coe et al. adjusted for the nesting of students within schools as well as pretest differences. While we also adjusted for such differences, we had to estimate an ICC, whereas Coe et al. were able to directly calculate one from the data at hand. As a result, we reran the analysis using the ES reported by Coe et al. (this situation of different effect sizes did not exist for the other three studies). This resulted in a statistically significant ($p < .001$) average weighted ES of .11 for the 6 + 1 Trait Writing program (confidence interval = 0.05 to .17). However, if we winsorized the sample size for the Coe et al. study as was done above, reducing its undue influence on this analysis, the average weighted ES remains at 0.11, but it is no longer greater than no effect ($p = .051$; confidence interval = −0.001 to .22).

## Discussion

### The Impact of Feedback on Students' Writing

As anticipated, classroom-based formative assessment that provided students with feedback on their written products or their progress in learning writing skills or strategies resulted in positive gains in children's writing. Such assessments resulted in almost two-thirds of a standard deviation gain in the quality of students' writing across 25 comparisons with students in grades 2 to 8. This exceeded the effects obtained for other writing treatments such as the process writing approach, sentence combining, teaching transcription skills, the use of word processing, and increasing

how much students write (see Graham, Harris, & Santangelo, 2015, in this issue). As an alternative reference point, the application of such formative assessment would move an average student (50th percentile on a measure of writing quality) to the 74th percentile.

Each of the four types of feedback tested in studies included in this meta-analysis also resulted in positive gains in the quality of students' writing. The largest effects were obtained for feedback from adults (seven-eighths of a *SD*), followed by self-feedback (sixth-tenths of a *SD*), peer feedback (slightly more than five-ninths of a *SD*), and computer feedback (three-eighths of a *SD*). The effects for each of these types of feedback were based on a small number of studies (the largest $N = 10$), however, and must be viewed as more tentative than the effect obtained for the total body of studies testing the effects of feedback to students on their written products or progress in learning. In addition, the magnitude of an effect for the four different types of feedback should not be interpreted to suggest that one type of feedback is more powerful than another, as these different forms of feedback were not directly compared in the studies reviewed here.

An important caveat in interpreting the findings for adult feedback is that such feedback took a variety of forms, ranging from teachers providing students with feedback on their writing, parents and other adults providing such feedback, and teachers making students aware of their progress in learning. It must further be noted that only one study examined the effects of teachers providing feedback on students' written text. It is surprising how few true or quasi-experiments tested this form of feedback, since this is one of the oldest and most common instructional procedures used by those who teach writing. More research is needed to test the effectiveness of teacher feedback on students' writing as well the effectiveness of other formative assessments used by teachers to provide feedback to students about their progress in learning to write.

Additionally, we were only able to locate four studies that tested the effectiveness of computers providing substantive feedback to students on their writing. While such feedback had a positive effect on the quality of students' writing, this finding must be interpreted even more cautiously than the findings for the other forms of feedback to students, as we loosened the criteria for study inclusion so that we would have at least four investigations. Consequently, the findings provide tentative support for the use of computer assessments by Smarter Balanced and PARCC, but additional research is needed to verify the effects obtained here. Moreover, we would encourage researchers to examine the interface between computer and teacher feedback to determine whether one form of assessment strengthens the impact of the other.

For all of the studies that examined feedback to students (adult, peer, computer, and self), we attempted to account for excess variability in effect sizes by examining specific study characteristics using meta-regression. We did not find that differences in magnitude of effects were statistically related to grade (elementary vs. middle school), type of feedback (structured vs. unstructured), or study quality. As additional studies examining feedback in writing become available, future meta-analyses need to return to this issue.

In summary, formative writing assessments where students obtained feedback about their writing or writing progress during the course of everyday classroom teaching and learning resulted in better student writing. Like Heritage (2010), we are

concerned that the strong reliance Smarter Balanced and PARCC are placing on summative and interim assessments represents a missed opportunity for these organizations to place more emphasis on the types of classroom-based formative assessments found to be effective in this review. We view this as unfortunate, as summative assessments have a spotty track record in the area of writing (Graham, Hebert, & Harris, 2011; Hillocks, 2002), and the impact of interim assessments is unproven (e.g., Goertz, Olah, & Riggan, 2009). Clearly, the assessment systems being developed by Smarter Balanced and PARCC would benefit by making formative classroom assessments that provide students with feedback a more integral part of their approach to improving children's writing.

## The Impact of Progress Monitoring on Students' Writing

Contrary to expectations, progress monitoring, as actualized in studies testing curriculum-based measurement, did not have a statistically significant impact across five investigations involving students in grades 2 to 8. The obtained effect was less than one-fifth of a standard deviation and not statistically significant. It was also smaller than the effect (0.25) used by What Works Clearinghouse (Graham, Bollinger, et al., 2012) to define an effect as substantially significant. An average student in the studies examined here would make a gain of seven percentile points. The relatively poor performance of this treatment was not a consequence of poor implementation or poor study quality, as studies were generally implemented with high fidelity and met most of the quality indicators.

This finding stands in contrast to the application of curriculum-based measurement studies in other academic domains, where this treatment produced sizable gains in academic achievement (e.g., average weighted ES = .70; Fuchs & Fuchs, 1986). This raises the question of why curriculum-based measurement did not produce a greater effect in this review of writing studies. One explanation for this involves the value of the measures used to monitor students' writing progress over time. Studies included in this review typically monitored correct spelling, correct word sequence, and total written words. While such measures can be reliably scored (see Table 11 in Graham, Harris, & Hebert, 2011), it is not clear how sensitive they are to changes in students' performance over short periods of time (such assessments are often given weekly). In addition, teachers may not be certain on how to parlay data on number of words written or correct word sequence into changes in how they teach.

In any event, there is a need for more research investigating the impact of curriculum-based measurement and other progress-monitoring approaches on students' writing performance. This will require identifying writing measures that are not only reliable, but sensitive to change in students' writing performance over a short period of time. It will further require identifying effective methods for helping teachers take the results of such assessments and translate them into productive methods of teaching.

## The Impact of the 6 + 1 Trait Writing Program on Students' Writing

The four studies testing the 6 + 1 Trait Writing program collectively produced a small average weighed effect for students in grades 3 to 6. While all of the effects were

positive, no single ES exceeded 0.19. No matter how we computed the average weighted ES for this treatment (i.e., winsorizing or not winsorizing sample size for Coe et al., 2011; using the ES we computed or the one reported by Coe et al.), it did not exceed 0.11. The only time that the average weighted ES was statistically significant was when we used the ES reported by Coe et al., and allowed it to exert an undue influence due to its large sample size. Even under this most favorable condition, an average student in the studies examined here would only make a gain of four percentile points as a result of participating in this program.

The relatively small effect for the 6 + 1 Trait Writing program obtained in this meta-analysis is likely to be disappointing to the many teachers across the United States who use this approach. One possible reason why the results were not stronger is that teachers did not apply the model as intended. This may have been the case in Coe et al. (2011), as the researchers indicated the "extent to which the model was actually implemented by treatment group teachers is unknown, as is the extent to which treatment group teachers implemented these strategies more than they were implemented by the control group teachers" (p. xiv). It is further possible that teachers needed more professional development and experience applying the model than the researchers offered in these studies (although considerable professional development was provided in at least one-half of the studies reviewed here; i.e., Arter, 1994, and Coe et al., 2011). Finally, the small effects obtained may be related to the quality of the studies testing this program. This explanation seems unlikely though as there was little difference in the ES obtained for stronger or weaker studies (see Table 1).

## Limitations

As with all meta-analyses, there are a number of limitations that need to be taken into account when interpreting the findings. First, meta-analyses, such as this one, involve aggregating findings from individual studies to draw general conclusions about one or more treatments or questions. The value of these conclusions depends on a variety of factors, such as the quality of the investigations and who participated in the studies, and must be interpreted accordingly.

Second, this review was limited to true and quasi-experiments (the only exception involved the inclusion of two studies that applied alternative designs to test computer-feedback effects). While the types of studies reviewed here control for a number of threats to internal validity, our decision to focus on these types of studies should in no way distract from the important contributions that other types of research (e.g., qualitative, single-subject) make to our understanding of the value of formative assessment procedures in writing.

When it was possible, we corrected for pretest difference when computing an ES by subtracting the pretest score from the posttest score for each condition. This was done to ensure, as much as possible, that the obtained ES was due to the treatment and not to initial difference between the treatment and control students. Such gain scores, however, are not without limitations, as some scholars claim they create problems of bias (e.g., overcorrect the pretest) and regression effects (e.g., Cook & Campbell, 1979). When pretests are not equivalent, the interpretation of a gain score may be problematic. This is more likely to be a problem with quasi-experiments where students are not randomly assigned. In true experiments, it is assumed that groups are equivalent, as randomization protects against regression toward the mean

and biased estimation by using a controlled design. We do not think that the use of gain or difference scores was problematic in this review for two reasons. One, most studies had equivalent scores at pretest (see Quality of Research section above). Two, for studies involving feedback (this included all but nine studies), there was no statistically significant difference between the average weighted ES for writing quality for true and quasi-experiments, $Q$ (between) = 1.19, $p$ = .28.

A final concern with meta-analysis involves the similarity of the outcomes and treatments in each study used to compute an average weighted effect size. As variability for each of these increases, the conclusions drawn become more clouded. We attempted to control for variability in treatments by analyzing specific formative assessment treatments separately. We attempted to limit variability in outcomes by only computing effect sizes for writing quality (although we expanded the permissible measures for CBM).

## Notes

The meta-analysis in this article was based, in part, on the meta-analysis presented in *Informing Writing: The Benefits of Formative Assessment* (Graham, Harris, & Hebert, 2011) commissioned and copyrighted by the Carnegie Corporation of New York. A free, downloadable copy of *Informing Writing* can be found on the Carnegie Corporation website at http://www.carnegie.org. Steve Graham is the Warner Professor in the Division of Educational Leadership and Innovation and Karen R. Harris is the Mary Emily Warner Professor in Mary Lou Fulton Teachers College at Arizona State University. Michael Hebert is assistant professor in the Department of Special Education and Communication Disorders at the University of Nebraska–Lincoln. Correspondence should be addressed to Steve Graham, Arizona State University, steve.graham@asu.edu.

## References

***References marked with an asterisk indicate studies included in the meta-analysis.**

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

*Adler, M. (1998). *The effects of instruction in six trait writing on third grade students' writing abilities and attitudes toward writing* (Unpublished master's thesis). Emporia State University, Emporia, KS.

Andrade, H. G., & 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., & 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.

Applebee, A., & Langer, J. (2006). *The state of writing instruction: What existing data tell us*. Albany, NY: Center on English Learning and Achievement.

*Arter, J. (1994). *The impact of training students to be self-assessors of writing*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Assessment Reform Group. (2002). *Assessment for learning: 10 principles*. Retrieved from https://castl.duq.edu/Conferences/Library03/PDF/Assessment/Ten_Principles.pdf

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

Bangert-Drowns, R., Kulik, C., Kulik, J., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research*, **61**, 213–238.

*Benson, N. (1979). *The effects of peer feedback during the writing process on writing performance, revision behavior, and attitude toward writing* (Unpublished doctoral dissertation). University of Colorado, Boulder.

Black, P., & Wiliam, D. (1988). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, **80**, 139–148.

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

*Caccamise, D., Franzke, M., Eckhoff, A., Kintsch, E., & 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.

*Coe, M., Hanita, M., Nishioka, V., & Smiley, R. (2011). *An investigation of the impact of the 6 +1 Trait Writing model on grade 5 student writing achievement* (NCEE 2012-4010). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Collopy, R. M. B. (2008). Professional development and student growth in writing. *Journal of Research in Childhood Education*, **23**, 163–178.

Common Core State Standards: National Governors Association and Council of Chief School Officers. (2010). Retrieved from http://www.corestandards.org/

Cook, T., & Campbell, D. (1979). *Quasi-experimental design and analysis for field settings*. Boston: Houghton Mifflin.

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

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

Culham, R. (2003). *6 + 1 Traits of Writing: The complete guide, grades 3 and up*. New York: Scholastic.

Deno, S. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children*, **52**, 219–232.

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

Fitzgerald, J., & Shanahan, T. (2000). Reading and writing relations and their development. *Educational Psychologist*, **35**, 39–50.

Formative Assessment for Students and Teachers, State Collaborative on Assessment and Student Standards. (2008, October). *Attributes of effective formative assessment*. Paper prepared for the Formative Assessment for Teachers and Students State Collaborative on Assessment and Student Standards of the Council of Chief State School Officers. Washington, DC: Council of Chief State School Officers.

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

Fuchs, L., & Fuchs, D. (1986). Effects of systematic formative evaluation on student achievement: A meta-analysis. *Exceptional Children*, **53**, 199–208.

*Fuchs, L. S., Fuchs, D., & 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., & 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., & Allinder, R. M. (1991b). The contribution of skills analysis to curriculum-based measurement in spelling. *Exceptional Children*, **57**, 443–448.

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

Goertz, M., Olah, L., & Riggan, M. (2009, December). *Can interim assessments be used for instructional change?* Philadephia: CPRE.

Graham, S., Bollinger, A., Booth Olson, C., D'Aoust, C., MacArthur, C., McCutchen, D., & Olinghouse, N. (2012). *Teaching writing in elementary school: A practice guide*. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Graham, S., Harris, K. R., & Hebert, M. (2011). *Informing writing: The benefits of formative assessment*. Washington, DC: Alliance for Excellence in Education.
Graham, S., Harris, K. R., & Santangelo, T. (2015). Research-based writing practices and the Common Core: Meta-analysis and meta-synthesis. *Elementary School Journal*, **115**, 498–522.
Graham, S., & Hebert, M. (2010). *Writing to reading: Evidence for how writing can improve reading*. Washington, DC: Alliance for Excellence in Education.
Graham, S., Hebert. M., & Harris, K. R. (2011). Throw em' out or make em' better? High-stakes writing assessments. *Focus on Exceptional Children*, **44**, 1–12.
Graham, S., McKeown, D., Kiuhara, S., & Harris, K. R. (2012). A meta-analysis of writing instruction for students in the elementary grades. *Journal of Educational Psychology*, **104**, 879–896.
Graham, S., & Perin, D. (2007a). *Writing Next: Effective strategies to improve writing of adolescent middle and high school*. Washington, DC: Alliance for Excellence in Education.
Graham, S., & Perin, D. (2007b). What we know, what we still need to know: Teaching adolescents to write. *Scientific Studies in Reading*, **11**, 313–336.
Griffin, B. (2000). *Quantitative research matrix*. Retrieved from http://coe.georgiasouthern.edu/foundations/bwgriffin/edur7130/quantitative_research_matrix.htm
*Guastello, E. F. (2001). Parents as partners: Improving children's writing. *Celebrating the Voices of Literacy: The Twenty-Third Yearbook of the College Reading Association: A Peer Reviewed Publication of the College Reading Association*, 279–293.
Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, **77**, 81–112.
Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin*, **92**, 490–499.
Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, **32**, 341–370.
Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, **29**, 60.
Heritage, M. (2010). *Formative assessment and next-generation assessment systems: Are we losing an opportunity*? Washington, DC: Council of Chief State School Officers.
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: Teachers College Press.
*Holliway, D. R. (2004). Through the eyes of my reader: A strategy for improving audience perspective in children's descriptive writing. *Journal of Research in Childhood Education*, **18**, 334–349.
*Holman, L. (2011). *Automated writing evaluation program's effects on student writing achievement* (Unpublished doctoral dissertation). Tennessee State University, Nashville.
*Jewell, J. (2003). *The utility of curriculum-based measurement writing indices for progress monitoring and intervention* (Unpublished doctoral dissertation). Northern Illinois University, Dekalb.
Kiuhara, S., Graham, S., & Hawken, L. (2009). Teaching writing to high school students: A national survey. *Journal of Educational Psychology*, **101**, 136–160.
*Kozlow, M., & Bellamy, P. (2004). *Experimental study on the impact of the 6 + 1 Trait Writing model on student achievement in writing*. Portland, OR: Northwest Regional Educational Laboratory.
Langer, J. (2011). *Envisioning knowledge: Building literacy in the academic disciplines*. New York: Teachers College Press.
Lipsey, M., & Wilson, D. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.
*Lumbelli, L., Paoletti, G., & 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., & Graham, S. (1991). Effects of a reciprocal peer revision strategy in special education classrooms. *Learning Disabilities Research*, **6**, 201–210.
McMaster, K. L., & Espin, C. A. (2007). Technical features of curriculum-based measurement in writing: A literature review. *Journal of Special Education*, **41**, 68–84.
*Meyer, E., Abrami, P., Wade, C., Aslan, O., & Deault, L. (2010). Improving literacy and metacognition with electronic portfolios: Teaching and learning with ePearl. *Computers & Education*, **55**, 84–91.

Morphy, P., & Graham, S. (2012). Word processing programs and weaker writers/readers: A meta-analysis of research findings. *Reading and Writing: An Interdisciplinary Journal*, **25**, 641–678.
National Center for Education Statistics. (2012). *The nation's report card: Writing 2011*. Washington, DC: Institute of Educational Sciences, U.S. Department of Education.
National Commission on Writing. (2003, April). *The neglected "R": The need for a writing revolution*. New York: College Board.
*Olson, V. L. B. (1990). The revising processes of sixth-grade writers with and without peer feedback. *Journal of Educational Research*, **84**, 22–29.
*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.
*Philippakos, Z. (2012). *Effects of reviewing on fourth- and fifth-grade students' persuasive writing and revising* (Unpublished doctoral dissertation). University of Delaware, Newark.
*Prater, D. L., & Bermudez, A. B. (1993). Using peer response groups with limited English proficient writers. *Bilingual Research journal*, **17**, 99–116.
*Reynolds, C. J., Hill, D. S., Swassing, R. H., & 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.
Rock, J. L. (2007). *The impact of short-term use of Criterion (SM) on writing skills in ninth grade* (No. RR-07-07). Princeton, NJ: ETS.
Rogers, L., & Graham, S. (2008). A meta-analysis of single subject design writing intervention research. *Journal of Educational Psychology*, **100**, 879–906.
*Rosenthal, 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., & Hogaboam-Gray, A. (1999). Effects of self-evaluation training on narrative writing. *Assessing Writing*, **6**, 107–132.
Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, **18**, 119–140.
*Schunk, D. H., & 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., & 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.
Stiggins, R. (2005). From formative assessment to assessment for learning: A path to success in standards-based schools. *Phi Delta Kappan*, **87**, 324–328.
Stinner, D., & Tennent, D. (2012). Losses to follow-up present risk to study validity: Differential attrition can be a shortcoming in clinical research. *AAOS Now* (February). Retrieved from http://www.aaos.org/news/aaosnow/feb12/research1.asp
Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.
*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.
*Wade-Stein, D., & Kintsch, E. (2004). Summary Street: Interactive computer support for writing. *Cognition and Instruction*, **22**, 333–362.
Wiliam, D. (2006). Formative assessment: Getting the focus right. *Educational Assessment*, **11**, 283–289.
*Wise, W. (1992). *The effects of revision instruction on eighth graders' persuasive writing* (Unpublished doctoral dissertation). University of Maryland, College Park.
Wolf, F. M. (1986). Meta-analysis: Quantitative methods for research synthesis. Newbury Park, CA: Sage.
*Wolter, D. R. (1975). *Effect of feedback on performance on a creative writing task* (Unpublished doctoral dissertation). University of Michigan, Ann Arbor.
*Young, E. A. (2000). *Enhancing student writing by teaching self-assessment strategies that incorporate the criteria of good writing* (Unpublished doctoral dissertation). Rutgers, the State University of New Jersey, New Brunswick, NJ.