VOLUME 44 NUMBER 4 DECEMBER 2011

FOCUS ON EXCEPTIONAL CHILDREN

It Is More Than Just the Message: Presentation Effects in Scoring Writing

Steve Graham, Karen R. Harris, and Michael Hebert

People write for many reasons. Writing is used as a tool to record ideas and information, communicate with others, chronicle experiences, express one's feelings, persuade others, facilitate learning, create imagined worlds, and evaluate students' competence (Graham, 2006). In some instances, the only intended reader of a piece of writing is the author. Examples of such writing include diaries, to do lists, and lecture notes. In other instances, writing is meant to be both read and formally evaluated by others. This kind of writing can range from term papers to state and federal writing assessments to writing requirements included as part of college entrance applications.

WRITING MATTERS

Many writing activities completed by students, including students with special needs, are not only evaluated by others but have specific consequences for the writer. For instance, evaluation of written homework activities, essay questions on exams, class papers, and so forth contribute to a student's eventual grade in a course. Writing tests administered by individual schools, states, or governments may be used to determine who graduates from high school, to identify students who need extra help with writing, and to evaluate the effectiveness of individual teachers and schools by determining how many students meet writing performance standards (Graham, Hebert, & Harris, 2011). Which college and even whether a student attends college may also be determined, at least in part, by a score on a standardized writing test.

The consequence of others' judgments about one's writing is not just limited to the classroom. Increasingly, employers make decisions about whom to hire and promote based on their evaluations of workers' writing competence, especially in white-collar jobs (National Commission on Writing, 2004, 2005). Even when writing is used for social purposes, others' judgments about one's writing can have real consequences. To illustrate, when the first two authors' daughter was a preteen she was a notoriously poor speller. One social consequence of this involved an online game, *Zena the Warrior Princess*, that she loved to play. Other online players often ignored her responses unless all of the words in her dialogue and messages were spelled correctly.

Students with disabilities often have difficulties with specific aspects of writing, like spelling (Graham & Harris, 2011), that may cloud others' judgments about the text they produce. It

Dr. Graham and Dr. Harris share the Currey Ingram Professor of Special Education and Literacy chair at Vanderbilt University. Mr. Hebert is a predoctoral fellow at Vanderbilt University.

is important, therefore, to determine what aspects of a writer's text influences others' judgments or evaluations about the quality of the message. This is important for both the writer and the reader. If writers, including students with special needs, are cognizant of the factors that shape others' evaluations, then they will presumably be more likely to attend to these factors while crafting and polishing their writing. Likewise, knowledge about the factors that shape as well as bias evaluations should lead to more valid assessments of writing by readers (including those who are tasked with the job of scoring students' writing), as such information increases the likelihood of focusing on the critical elements of good writing while minimizing the potential biasing effects of more trivial features like spelling miscues.

PRESENTATION EFFECTS: SPELLING, HANDWRITING, AND GRAMMAR

The present article focuses on a set of factors, referred to here as *presentation effects*, which may influence readers' or scorers' judgments about written text. This includes text written by students with and without disabilities. We illustrate one aspect of the presentation effect with a story about an attempted robbery in Corpus Christi, Texas (Lederer, 2000). While standing in line, the robber wrote a stickup note on a Bank of America deposit slip but became worried that someone might have seen him write the note. As a result, he went across the street to Wells Fargo and handed the teller his note which read: "This iz a stickup. Put all your muny in this bag." Based on his spelling, the teller assumed that the robber was not very bright and told him she could not accept a stickup note written on a Bank of America slip. She sent him back across the street, where he was subsequently arrested. While the message in the robber's note was clear and appropriate, the presentation effect of misspelled words influenced the readers' judgments about the message and the person who wrote it.

The earliest and most frequently studied presentation effect involved the influence of text legibility in scoring writing. In a 1927 study, James quoted a hypothetical student who asked whether he may make a neater copy of his paper so that he will receive a grade of B instead of C. His experiment provided support for the student's contention. James asked teachers to twice score the writing quality of four average compositions written by high school seniors. Two of these average compositions had good handwriting, whereas two had poor handwriting. On the second scoring (which took place 2 months later to minimize teachers' remembrance of their initial score for each paper), the legibility of two papers was reversed: good to poor and poor to good. James found that legibility resulted in a 9-point difference in students' scores on papers of similar writing quality, with more legible papers receiving higher scores and less legible ones lower scores. This caused him to question what was "the chief factor in determining the teacher's grade on a theme: the legibility of the handwriting or the quality of the composition" (p. 180). He further argued that the influence of legibility on scoring the quality of students' writing was too large.

A slightly different concern about presentation effects was raised by Scannell and Marshal (1966). They noted that writing is often used to assess a student's knowledge about content. For example, students may be asked to demonstrate their knowledge about a topic on an essay exam, through a written paper, or on a standardized test. They argued that the scores assigned to such text may reflect more than the quality of the written answer. Such scores may be biased by presentation effects due to either spelling or grammar errors, as these flaws are readily apparent to scorers who are likely to be differentially sensitive to them. As Markham (1976) indicated a decade later, noncontent factors (e.g., handwriting and spelling and grammar errors) may exert too much influence on the scores assigned to text designed to demonstrate students' content knowledge.

FOCUS ON EXCEPTIONAL CHILDREN

ISSN 0015-511X

FOCUS ON EXCEPTIONAL CHILDREN (USPS 203-360) is published monthly except June, July, and August as a service to teachers, special educators, curriculum specialists, administrators, and those concerned with the special education of exceptional children. This publication is annotated and indexed by the ERIC Clearinghouse on Handicapped and Gifted Children for publication in the monthly *Current Index to Journals in Education* (CIJE) and the quarterly index, *Exceptional Children Education Resources* (ECER). The full text of *Focus on Exceptional Children* is also available in the electronic versions of the *Education Index*. It is also available in microfilm from Serials Acquisitions, National Archive Publishing Company, P.O. Box 998, Ann Arbor, MI 48106-0998. Subscription rates: individual, $50 per year; institutions, $68 per year. Printed in the United States of America. Periodical postage is paid at Denver, Colorado. **POSTMASTER:** Send address changes to:

Love Publishing Company
Executive and Editorial Office
P.O. Box 22353
Denver, Colorado 80222
Telephone (303) 221-7333

It is not too difficult to imagine how noncontent factors such as poor handwriting or spelling or grammar errors influence the scoring of students' writing. As early as 1929, Sheppard described three possible reactions of a scorer to such flaws. One, the scorer reads part of the text with such flaws but puts it aside before finishing it and gives it a low grade. Two, the scorer realizes at a glance that such flaws are present and does not read it, giving the paper the grade he or she thinks the student deserves. Three, the scorer reads the paper despite one or more of these flaws, but judgments about the quality of the student's response are diminished. In any event, the score assigned may not be representative of the quality of the content included in the paper.

For students with disabilities, presentation effects may be especially troublesome. The papers these students produce are often difficult to read because of poor handwriting and typically contain many spelling and grammatical errors (Graham & Harris, 2011). If presentation effects are in fact real, this has serious consequences for the grades they receive on written classroom assignments and for their writing achievement on high-stakes tests used by districts, states, and the federal government. Of course, such presentation effects have consequences for typically developing students, too! For example, students' handwriting often becomes less legible across the grades as this skill becomes more fluent (Graham & Weintraub, 1996), and typically developing students continue to make spelling and grammar errors in their writing even when they are in college.

PRESENTATION EFFECTS: WORD PROCESSING PRINTED TEXT

In addition to possible presentation effects due to handwriting, spelling, and grammar, the advent and wide use of word processing introduced another possible noncontent biasing factor into the scoring of writing. During the 1990s, researchers began expressing concern that word processing printed text may be scored more harshly than handwritten text. In a study by Arnold et al. (1990), reported by Powers, Fowles, Farnum, and Ramsey (1994) scorers indicated they preferred scoring handwritten text over word-processing printed text, even though the former was harder to read. The scorers also appeared to empathize more with the writer of handwritten text, giving them benefit of the doubt when they encountered difficulty in reading handwritten text and even going so far as to mentally transform or fill in perceived gaps in the text when scoring it. Finally, Arnold and his colleagues suggested that scorers have higher expectations for word-processing produced text than for handwritten text, expecting fully polished and edited products and penalizing such text if this was not the case.

We currently live in a world where students, including students with disabilities, produce both handwritten and word-processing text for school. In the United States, for example, students in grades 1–12 still produce most of their text for class by hand (Cutler & Graham, 2008; Gilbert & Graham, 2010; Kiuhara, Graham, & Hawken, 2009), but more than one half of high school teachers now indicate that their students complete a writing assignment via word processing at least once a month. Likewise, most high-stakes writing assessments required by schools, states, and the federal government in the U.S. are still handwritten. However, this is also starting to change, as prominent high-stakes writing tests, such as those administered by the National Assessment of Educational Progress, have adopted word-processing administered writing tests. This trend of word-processing produced writing assessment is likely to quicken, as both of the common assessment consortia in the U.S. (Smater and PARCC) are currently developing formative and summative assessments, including electronic ones (Gewertz & Robelen, 2010), for the Common Core State Standards Initiative (http://www.corestandards.org/the-standards/english-language-arts-standards). Therefore, if word-processing printed text is indeed scored more harshly than handwritten text, validly scoring writing across writers (e.g., those who write via word processing and those who do not) and within writers (e.g., students who use both modes of writing) becomes a challenge. This is an issue for students with and without disabilities.

PURPOSE OF THIS ARTICLE

The purpose of this review is to determine the impact of handwriting, spelling, grammar, and word-processing printed text on the scoring of students' writing (both school-age and college students). Almost all of the studies in this article employed the same basic format for testing the biasing effects of each of these noncontent presentation factors. In its simplest form, this involved modifying a paper written by a student to create two or more versions of the paper containing the exact same content but differing in terms of legibility, number of spelling errors, number of grammar errors, or whether it was handwritten or word-processing printed (this always involved creating a second word-processing version of a handwritten paper). The content or quality of papers was then scored by teachers, teachers in training, graduate students, or college instructors.

The review method applied in this article was meta-analysis. This systematic approach to reviewing the literature is used to summarize the direction and magnitude of the effects obtained in a set of empirical research studies examining relevant investigations (Lipsey & Wilson, 2001). Meta-analysis was well suited to the goals of this review, which involved determining whether each of the four presentation factors did in fact bias writing assessments, and, if so, how strong was the impact of each factor. This is an especially important goal, as the elimination of presentation effects, from both

classroom and high stakes writing assessments for students with and without disabilities, is likely to be time consuming and expensive. Thus, there is no reason to undertake corrective action if the effects are modest to nonexistent.

Based on our review of literature concerning the biasing effects of the four presentation factors examined in this review, we expected that a less legible version of text would receive a lower writing score than a more legible version of the same text, that both spelling and grammar errors would lower writing scores, and that word-processing printed text would be scored more harshly than handwritten text.

METHODS OF THE REVIEW

Before presenting the findings of the review, we first present the methods for locating, including, and coding studies as well as the methods for calculating effect sizes..

Study Inclusion/Exclusion Criteria

Studies had to meet the following five criteria to be included in this meta-analysis. The study (a) assessed the effects of handwriting, spelling, grammar, or word-processing printed text in scoring students' writing; (b) involved an experimental design (true-experiment, quasi-experiment, or within-participant design; (c) involved the scoring of text produced by students in grades 1 through college; (d) was presented in English; and (e) contained the statistics necessary to compute a weighted effect size (or statistics were obtained from the authors).

For handwriting, spelling, and grammar presentation effects, studies had to employ a paradigm where the writing of different versions of the same text (e.g., more or less legible versions of the same text) was scored by teachers, teachers in training, graduate assistants, and college instructors. This typically involved taking a single text and modifying it some way (e.g., eliminating all spelling errors and adding a specific number of spelling errors). In addition, more than one text could be modified (e.g., Russell & Tao, 2004a, modified 40 student compositions). We decided to include a single study that employed a slightly different paradigm. Klein and Taub (2005) took four papers written by sixth-grade students and had two of the papers copied verbatim by a student with poor handwriting. The other two papers were copied verbatim by a student with good handwriting. Because the four papers were similar in overall writing quality (established by three judges), we included this study in our review.

Studies that examined the effects of word-processing printed text employed a slightly different paradigm. In these studies, the handwritten text of a range of student writers were converted to word-processing printed text retaining all errors. These compositions were then scored by teachers, teachers in training, graduate students, or college instructors.

Search Strategies Used to Locate Studies

A search that was as broad as possible was undertaken to identify relevant studies for this review. In October, 2010, electronic searches were run in multiple databases, including ERIC, PsychINFO, ProQuest, Education Abstracts, and Dissertation Abstracts, to identify possible studies. Descriptors included assessment, evaluation, handwriting and writing quality, spelling and writing quality, grammar and writing quality, word processing, portfolio assessment, performance assessment, curriculum-based assessment, automated essay scoring, computer scoring, analytic quality, holistic quality, high-stakes assessment, and state writing assessments.

These electronic searches identified close to 7,000 possible items. The title and abstract for each entry was read by the first author of this review (if an abstract was not available, the title was examined). If an item looked promising, it was obtained. Furthermore, we conducted a hand search of the following journals: *Assessing Writing, Research in the Teaching of English,* and *Written Communication.* Once a document was obtained, the reference list was examined in order to identify additional promising studies. Of 55 documents collected, we found 17 papers that contained experiments that met the inclusion criteria. Thirty-two papers were eliminated because they did not directly assess the effects of handwriting, spelling, grammar, or word-processing printed text in scoring students' writing. Five papers that assessed such effects were eliminated because they did not provide the needed statistics for computing a weighted effect size.

Categorizing Studies into Presentation Effect Conditions

Each study was read by the first author and placed (if possible) into one of the four presentation effects categories: handwriting, spelling, grammar, or word-processing printed text. Studies that did not fit neatly within one of these four categories were held apart until all studies were read and sorted. At this point, the studies placed in each of the four categories were reread to confirm their initial placement. This process did not result in the creation of any additional categories, nor did it result in changing the initial placement of any study.

Coding of Study Features

Each study was coded for particular study characteristics and to determine whether specific indicators of study quality were met. Study characteristics included design of study, grade of students who produced the text, type of text (e.g., narrative), number of texts scored, modification of text, type of scorer (and whether they received training), number of scorers, writing outcome (i.e., quality of writing and quality of content), study quality (percentage of quality indicators met by a study), and publication type.

There were five quality indicators: (1) design (true-experiment, quasi-experiment, and within-participant design); (2)

total attrition was less than 10% of total sample; (3) total attrition was less than 10%, and equal attrition across conditions was evident (i.e., conditions did not differ by more than 5%); (4) Hawthorne effects were not evident (i.e., researchers put into place conditions for controlling for Hawthorne effects, such as an alternative condition that controlled for attention and time); and (5) ceiling/floor effects for the writing assessment were not evident (more than 1 *SD* from floor and ceiling) . Each quality indicator was scored as 1 (met) or 0 (not met). The only exception involved design, where a 1 was assigned if the study was a true experiment; 0.5 was assigned if it was a quasi-experiment, and 0 was assigned if it was a within-participant design. A total score was calculated for each study. This was converted to a percentage by dividing the obtained score by total possible points allowable and multiplying by 100%.

Coding for study descriptors and quality indicators were independently scored a second time to establish reliability. Agreement was 98% across all categories.

Calculation of Effect Sizes

An effect size (*ES*) was calculated by subtracting the mean score of the treatment group (type of scorer, type of student who wrote the different versions of the paper scored, type of writing instrument [pen or pencil] used in creating the scored paper, and type of text [word-processing text typed in manuscript and cursive or word-processing text single and double spaced]). To aggregate data, we applied the procedures recommended by Nouri and Greenberg (Cortina & Nouri, 2000). 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:

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

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:

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

Average weighted ES

This meta-analysis used a weighted random-effects model. For each presentation factor (e.g., handwriting), we calculated an average weighted *ES* (weighted to take into account sample size by multiplying each *ES* by its inverse variance). For studies examining handwriting, spelling, or grammar effects, sample size was based on number of scorers (many of these studies involved different versions of a single composition that was scored by multiple scorers). For presentation effects due to word-processing printed text, sample size was based on number of samples scored (in this case a relatively large number of papers were scored by a small number of scorers).

We made an a priori decision to calculate an average weighted *ES* for a presentation factor (e.g., spelling) only if there were four or more studies testing it. The precedence for this decision was that this was the smallest number of ESs included in any writing treatment analyzed by Hillocks (1986) in his seminal review of the writing intervention literature and Graham and Perin (2007) in their review 20 years later. The 17 studies yielded 9 effects for handwriting, 7 effects for word-processing printed text, 5 effects for spelling, and 5 effects for grammar. As a result, we calculated an average weighted *ES* for each of these presentation factors.

In addition to calculating an average weighted *ES* for each presentation factor (e.g., grammar), the confidence interval and statistical significance of the obtained weighted *ES* were also calculated as were two measures of homogeneity (Q and I^2). The homogeneity measures allowed us to determine whether variability in the *ES*s for a presentation factor (e.g., word processing printed text) was larger than expected based on sampling error alone.

If homogeneity in *ES*s for a specific presentation factor exceeded sampling error alone and that factor had at least 8 *ES*s, we conducted moderator analysis to determine whether identifiable differences between studies could account for this excess variability. Only one presentation factor, handwriting, included enough *ES*s to conduct such an analysis. Because of the small number of effects, we examined only a single moderator variable: text produced by grade 1–12 students versus text produced by college students, assuming that scorers would have different expectations for the handwriting of younger versus older and more polished students. This moderator analysis involved an analogue similar to a one-way ANOVA (Hedges & Olkin, 1985). Handwriting *ES*s were grouped into two mutually exclusive categories (i.e., *ES*s for text produced by school-age students and *ES*s for text produced by college students), and the homogeneity of effect sizes within each category was tested as was the difference between the levels of the two mutually exclusive categories.

Maintaining Statistical Independence

To avoid inflating the sample size (Wolf, 1986), only one ES per paper was applied when computing an average weighted ES for each of the four presentation factors. This also had the added benefit of ensuring that we did not violate the assumption of independence of data underlying moderator analyses that was undertaken with handwriting.

There were two exceptions to this rule. Sheppard (1929) included two experiments in his paper. In both experiments the same two versions of a writing sample (more or less legible) were scored, but a different set of scorers were employed in each experiment. Russell and Tao's (2004b) paper included what we viewed as three experiments. One of the experiments involved teachers scoring the same text produced by grade 4 students in either a handwritten or word-processing form. The other two experiments involved similar scoring of text produced by grade 8 and 12 students, respectively. Thus, Sheppard (1929) contributed 2 *ES*s to the calculation of an average weighted *ES* for handwriting, whereas Russell and Tao (2004b) contributed 3 *ES*s to our examination of the biasing effects of word-processing printed text.

RESULTS

Table 1 contains information on the studies testing the effects of each of the presentation factors. Some studies are included under more than one presentation factor (as they tested more than one presentation effect), and presentation factors are presented in the same order they were initially introduced. The following information was presented about each study: reference, design (true-experiment, quasi-experiment, within-participant design), grade of students who produced the text, type of text (e.g., narrative), number of samples scored, how the text was modified, type of scorer (including whether they were trained to score writing samples), number of scorers, writing outcome (writing quality versus writing content), study quality (percentage of quality indicators met by a study), and *ES*. Studies are arranged

TABLE 1.
Information on Individual Studies for Each Presentation Factor

Study	Design	Ss	Text	Samples	Modification	Scorer	N Scorer	Out-come	Quality	*ES*
Handwriting										
Klein & Taub, 2005	WP	6	PN	4	PHW Ss with PHW rewrote GHW Ss with GHW rewrote	T (NTr)	53	W	80%	-1.10
Sheppard, 1929 (Exp. 1)	WP	8	E	2	PHW = 30 on 100 pt scale GHW = 90 on 100 pt scale	T (NTr)	225	NS	80%	-1.10
Sheppard, 1929 (Exp. 2)	WP	8	E	2	PHW = 30 on 100 pt scale GHW = 90 on 100 pt scale	T (NTr)	225	NS	80%	-1.30
Soloff, 1973	QE	11	E	2	PHW Ss with PHW rewrote GHW Ss with GHW rewrote	T (NTr)	32	C	90%	-0.94
Marshall & Powers, 1969	TE	12	E	1	PHW Ss with PHW rewrote GHW Ss with GHW rewrote	TIT (NTr)	70	C	100%	-0.38
Huck & Bounds, 1972	QE	U	E	2	PHW Ss with PHW rewrote GHW Ss with GHW rewrote	TIT (NTr)	34	C	100%	-0.45
Chase, 1968	QE	U	E	2	PHW = 23 on 100 pt scale GHW = 88 on 100 pt scale	TIT (NTr)	64	C	90%	-0.70
Chase, 1979	QE	U	E	2	PHW at lower end of Ayres GHW at upper end of Ayres	TIT (NTr)	62	C	90%	0.07
Bull & Stevens, 1979	QE	U	E	1	PHW which was legible GHW neatly written	T (NTr)	48	W	100%	-0.54
Spelling										
Russell & Tao, 2004a	QE	8	ES	40	PS = Ss errors retained GS = 0% of words misspelled	T (Tr)	8	W	90%	-0.32
Scannell & Marshall, 1966	TE	12	E	1	PS = 3% of words misspelled GS = 0% of words misspelled	TIT (NTr)	66	C	100%	-0.59
Marshall, 1967	TE	12	E	1	PS = 5% of words misspelled GS = 0% of words misspelled	T (NTr)	150	C	100%	-0.50
Marshall & Powers, 1969	TE	12	E	1	PS = 5% of words misspelled GS = 0% of words misspelled	TIT (NTr)	70	C	100%	-0.32
Chase, 1968	QE	U	E	2	PS = 13% of words misspelled GS = 0% of words misspelled	TIT (NTr)	64	C	90%	0.03

TABLE 1. *(continued)*

Study	Design	Ss	Text	Samples	Modification	Scorer	N Scorer	Out-come	Quality	*ES*
Grammar										
Yeh, 1998	WP	7	P	8	PG = incorrect conventions GG = correct conventions	T (NTr)	8	W	80%	-0.77
Scannell & Marshall, 1966	TE	12	E	1	PG = 10 grammar errors GG = no grammar errors	TIT (NTr)	66	C	100%	-0.48
Marshall, 1967	TE	12	E	1	PG = 18 grammar errors GG = no grammar errors	T (NTr)	150	C	100%	-0.49
Marshall & Powers, 1969	TE	12	E	1	PG = 18 grammar errors GG = no grammar errors	TIT (NTr)	70	C	100%	-1.00
Freedman, 1979	TE	U	E	8	PG = weaker form GG = stronger form	CI (Tr)	12	W	100%	-0.24
Word Processing Printed Text										
Russell & Tao, 2004b	QE	4	ES	52	HW papers typed with errors	T (Tr)	6	W	90%	-0.64
Russell & Tao, 2004b	QE	8	ES	60	HW papers typed with errors	T (Tr)	6	W	90%	-0.83
Russell & Tao, 2004a	QE	8	ES	40	HW papers typed with errors	T (Tr)	8	W	90%	-0.47
Wolf, Bolton, Feltovich, & Welch, 1993,* (Study 2)	TE	10	ES	80	HW papers typed with errors	T (NTr)	18	W	100%	-0.27
Russell & Tao, 2004b	QE	12	ES	60	HW papers typed with errors	T (Tr)	6	W	90%	-0.55
Sweedler-Brown, 1991	TE	U	ES	61	HW papers typed with errors	CI/GR (Tr)	28	W	100%	-0.24
Powers, Fowles, Farnum, & Ramsey, 1994	QE	U	PN, P	32	HW papers typed with errors	CI, T (Tr)	4	W	90%	-0.50

Note. WP = Within Participant; QE = Quasi-Experiment; TE = True Experiment; Ss = Student Grade; U = University; PN = Personal Narrative, N = Narrative; ES = Essay; P = Persuasive; PHW = Poor Handwriting; GHW = Good Handwriting; PS = Poor Spelling; GS = Good Spelling; PG = Poor Grammar; GG = Good Grammar; HW = Handwritten; T = Teachers; TIT = Teachers in Training; NTr = No Training; Tr = Training; CI = College Instructors; GR = Graduate Students; W = Writing; NS = Not Specified; C = Content

under each presentation factor from earlier to later grades for the students who wrote the compositions. Table 2 includes the number of studies, average weighted *ES*, confidence interval, standard error, and statistical significance for each presentation factor as well the two heterogeneity measures (Q and I^2).

Quality of Studies

In terms of the five quality indicators that were assessed, the overall quality of research assessing the four presentation factors was generally strong. All studies met at least four of the five quality indicators. Hawthorne effects, issues involving subject attrition, and ceiling/floor problems with the writing outcome measure were not evident in any study. While we categorized less than one third of the studies as true experiments, it is possible that some of the studies categorized as quasi-experiments did in fact involve randomization. In almost all of the quasi-experiments, the authors did not indicate how subjects were assigned to conditions.

Handwriting Presentation Effects

We calculated 9 *ES*s from 8 papers (Sheppard, 1929, contained two experiments) that tested the effects of variations

TABLE 2:
Average Weighted Effect Sizes and Confidence Intervals for Presentation Factors

				Test of Null Hypothesis		Heterogeneity	
Presentation Factor	Studies	Effect Size	Confidence Interval	Standard Error	*p*-value	*Q*-Value	I^2
Handwriting Legibility	9	-0.76	(-1.05, -0.50)	.151	< .001	**41.60	80.77
School-Age Text	5	-1.03	(-1.29, -0.76)	.135	< .001	*12.94	69.09
College Text	4	-0.39	(-0.75, -0.04)	.183	.03	5.14	41.53
Spelling Errors	5	-0.38	(-0.56, -0.20)	.092	< .001	4.50	11.03
Grammar Errors	5	-0.56	(-0.77, -0.36)	.105	< .001	4.18	4.35
Word Processing Printed Text	7	-0.48	(-0.64, -0.32)	.081	< .001	7.40	18.97

Note. * p < .05
** p < .01

in handwriting legibility in scoring students' writing (see Table 1). In all but one of these investigations, handwriting legibility was manipulated by creating more and less legible versions of a paper (in most instances two papers were modified in this way). These modifications were created by the researcher or by having students with good and poor handwriting rewrite a specific paper. It is important to note that the compositions that were scored in these studies were generally produced by average to slightly above-average writers, and while legibility was modified to be poor or good, it was not modified so that it represented the worst or best possible versions of legibility. The one exception involved a study by Klein and Taub (2005). They did not create multiple samples of a student's paper that differed only in legibility; rather, they identified samples of students' writing that were similar in terms of writing quality and had a student with poor handwriting and one with good handwriting recopy half of the papers to be scored by teachers.

The text that was scored by teachers or teachers in training was created by students in grades 6 through college (see Table 1). It was mostly expository text that students created to display their knowledge, and, as a result, papers were primarily scored in terms of quality of content. The one clear exception to this was Klein and Taub (2005), where teachers were directed to score overall quality of writing. Scorers were not taught how to score compositions in any of the studies, although 56% of the experiments involved experienced teachers. Only one study was categorized as a true experiment, with the rest identified as either a quasi-experiment or within-participant design.

Student papers were scored more harshly by teachers and teachers in training when an identical composition (or a composition of similar writing quality) was less legible versus more legible. All but one study produced a negative effect for less legible handwriting, yielding a statistically significant average weighted *ES* of -0.76. The *Q* test for heterogeneity was statistically significant, however, and I^2 indicated that 81% of the variance was due to between-study factors (see Table 2).

To determine whether excess variability was due to grade of the writer (school-age versus college), we conducted a moderator analysis. The average weighted *ES* for school-age writers (-1.03) was statistically larger than the average weighted *ES* for college writers (-0.39), $Q\ (between) = 10.33$, $p = .001$. The average weighted effect size was statistically greater than no effect for both school-age and college writers (see Table 2), and type of writer accounted for some of the excess variance, as most of the excess variance in *ES*s for college writers was accounted for by sampling error alone. This was not the case for school-age writers though.

Spelling Presentation Effects

We calculated 5 *ES*s that tested the effects of variations in spelling errors in scoring students' writing (see Table 1). The good spelling condition in all 5 investigations involved no errors, whereas the poor spelling condition ranged from 3% of words misspelled to 13% of words misspelled (we were not able to determine the exact percentage of words misspelled in Russell & Tao, 2004a). As was the case with handwriting, the compositions were typically average papers, although the Russell and Tao (2004a) investigation involved papers that represented a broader range of writing quality.

The paper(s) that were scored by teachers or teachers in training was created by students in grades 8 through college (see Table 1). It was mostly expository text that students created to display their knowledge, and, as a result, papers were primarily scored in terms of quality of content. Raters did

not receive training on how to score compositions in these studies. The one exception was the study by Russell and Tao (2004a), where the text was an essay written as part of a state writing test, and teachers received training in how to score quality of writing (the outcome measure in this investigation). Sixty percent of the studies were true experiments; the rest of the investigations were quasi-experiments.

A paper with spelling errors was scored more harshly by teachers and teachers in training than the exact same paper with no spelling errors. All but one study produced a negative effect for papers with spelling errors, yielding a statistically significant average weighted *ES* of -0.38. The *Q* test for heterogeneity was not statistically significant, and most of the variance in *ES*s was accounted for by sampling error alone (see I^2 statistic in Table 2).

Grammar Presentation Effects

We calculated 5 *ES*s that tested the effects of variations in grammar errors in scoring students' writing (see Table 1). The good grammar condition for a paper mostly involved no grammar errors (Freedman, 1979, created a version of a paper that had fewer grammar errors), whereas the poor grammar condition included 10 to 18 errors, or it was difficult to determine the exact number of errors (Freedman, 1979; Yeh, 1998). Again, papers were mostly typical compositions produced by students in grades 7 through college.

Papers were scored by teachers of middle and high school students, teachers in training, and college instructors (see Table 1). The scored text was mostly expository (one study involved persuasive writing; Yeh, 1998). In three of the studies, raters were directed to score papers for quality of content, whereas papers were scored for writing quality in the other 2 investigations. Scorers were only trained in one study (Freedman, 1979), and all but one study was categorized as a true experiment (Yeh, 1998, used a within-participant design).

A paper with more grammar errors was scored more harshly by teachers, teachers in training, and college instructors than the exact same paper with no or fewer grammar errors. All of the studies produced a negative effect for papers with more grammar errors, resulting in a statistically significant average weighted *ES* of -0.56. The *Q* test for heterogeneity was not statistically significant, and most of the variance in *ES*s was accounted for by sampling error alone (see I^2 statistic in Table 2).

Word-Processing Printed Text

We calculated 7 *ES*s from 5 papers (Russell & Tao, 2004b, yielded 3 *ES*s) that tested whether word-processing printed text was judged more harshly than handwritten text (see Table 1). Handwritten text produced by students in grades 4 to college was converted to identical word-processing printed text in each study. Students' text generally represented a broad range of writing ability, and in most studies students produced an essay as part of a state writing test. In all cases, teachers, teachers in training, graduate students, and college instructors scored students' compositions for writing quality after they were trained to apply the scoring procedures. The only exception involved Wolf, Bolton, Feltovich, and Welch (1993), where no training was provided. Five of the seven comparisons were from quasi-experiments (the other 2 were tested with a true experiment).

A paper printed in word-processing text was scored more harshly by teachers, teachers in training, graduate students, and college instructors than the exact same paper when it was handwritten. All of the studies produced a negative effect for word-processing printed text, resulting in a statistically significant average weighted *ES* of -0.48. The *Q* test for heterogeneity was not statistically significant, and most of the variance in *ES*s was accounted for by sampling error alone (see I^2 statistic in Table 2).

DISCUSSION

The writing of school-age and college students, those with and without disabilities, is frequently graded and scored by others. Sometimes this involves scoring students' writing to determine how well a particular piece of text is written (e.g., grading an assigned paper). At other times, it involves scoring one or more pieces of writing to gauge how well students write in general (e.g., norm-reference standardized writing tests, state writing assessments, entrance exams for college). At still other times, writing is scored to determine what a student knows about a particular subject (e.g., essay exam in a history class). Such assessments occur both in the classroom and outside of it and have consequences for students, teachers, and schools.

A long-standing concern in scoring the writing of students with and without special needs for these various purposes is that noncontent factors (e.g., handwriting, spelling errors) exert undue influence on the outcomes of such assessments. More than 80 years ago, James (1927) cautioned that noncontent factors like text legibility exert too much influence on the grades teachers assign to a particular piece of writing. Forty years later, Scannell and Marshal (1966) questioned what was assessed when writing is used to evaluate students' knowledge, as teachers are readily aware of and influenced by noncontent flaws (e.g., spelling errors) in youngsters' answers. Such concerns are especially problematic for students with special needs, as their handwriting is often poor, and spelling and grammar errors are common in their writing (Graham & Harris, 2011). The current meta-analysis provides a long overdue systematic review of the impact of four noncontent presentation factors (handwriting, spelling, grammar, and word processing printed text) on the scoring of students' written text.

Caveats

Before summarizing the findings from this meta-analysis, it is important to consider five factors that can influence interpretation. One, this review involved aggregating the findings from individual studies to draw conclusions about specific presentation effects (e.g., handwriting). The value of any conclusion drawn depends on the quality of the investigations testing the effect. Fortunately, the studies included in this review were generally well designed. Problems with attrition, Hawthorne effects, and ceiling and floor issues with the writing outcome measure were not evident in any of the studies. The one area of concern involved the relatively small percentage (29%) of true experiments in this body of research. Even so, the true experiments in this review produced findings similar to those yielded by quasi-experiments and within-participant studies (average weighted *ES* across all findings for true experiments was -0.44 versus -0.49 for quasi-experiments and within-participant studies).

An important issue involving meta-analysis is the comparability of outcome measures on which the effect sizes are based. The approaches used to evaluate writing in the studies in this review were varied (and sometimes poorly described). This introduces unwanted noise into the machinery of our meta-analysis and must be considered when interpreting the findings. Moreover, in studies examining handwriting, spelling, and grammar presentation effects, the primary means of scoring text centered on assessing the quality of the content in students' papers, whereas studies examining whether word-processing printed text was scored more harshly assessed just writing quality. Additional research is needed to determine whether both quality of content and writing are equally influenced by each of the presentation factors examined here.

None of the presentation factors examined in this review have been studied extensively. Even for the most studied effect, handwriting legibility, we were able to compute only 9 effects. It is also important to note that we did not examine all possible noncontent presentation factors (e.g., type of word-processing font). Clearly, additional research is needed.

It must further be noted that studies examining presentation effects due to handwriting, spelling, and grammar did not establish at what specific point poor legibility or number of errors start to bias the assessment of writing. The available studies did not, for example, examine whether presentation effects were evident when just 1% or 2% of words were misspelled. Thus, the current body of research in this area is not developed enough to indicate exactly when biasing is evident.

Finally, our search strategy was comprehensive and designed to identify all published and unpublished studies. However, we were not able to locate one of the unpublished studies identified in our search (Arnold et al., 1990), and only one of the 17 papers included in this review was unpublished. This raises the possibility that there may be other studies, both published and unpublished, that we were unable to locate.

Handwriting, Spelling, Grammar, and Word-Processing Printed Text

As predicted, all four of the presentation factors examined in this meta-analysis influenced the evaluation of students' writing. These effects ranged from moderate to large, depending upon the presentation factor and type of student, with spelling producing the smallest average weighted effect (three eighths of a standard deviation) and handwriting legibility with school-age students the largest (a full standard deviation). To place the obtained effects in perspective, the score for a typical paper would drop from the 50th percentile to between the 22nd and 10th percentiles (95 out of 100 times) if it was written by a school-age student with poor but readable handwriting (which is characteristic of many students with special needs; Graham & Weintraub, 1996). The drop for a college student with poor handwriting would be less dramatic, as a paper at the 50th percentile would fall to between the 48th and 23rd percentiles. Similarly, spelling and grammar errors (which are quite common in papers written by students with disabilities; Graham & Harris, 2011) had a deleterious effect on writing scores, as such miscues would drop a paper from the 50th percentile to between the 42nd and 29th percentiles and the 36th to the 22nd percentiles, respectively. When a paper printed from text produced on a word processor is compared to the same paper written by hand, writing scores would fall from the 50th percentile to between the 37th and 26th percentiles. Of course, writing scores would improve equally in the opposite direction for students with good handwriting, little to no spelling or grammar errors, or handwritten text when compared to word-processing printed text.

What We Still Need to Know

The single moderator analysis we were able to undertake revealed that presentation effects for handwriting were stronger for younger students (school age) than older ones (college). Although the effects of spelling and grammar on college students' writing scores was limited to one study each, they were consistent with the findings for handwriting. Chase (1968) obtained a positive *ES* (0.03) when spelling errors were included in text written by a college student, whereas Freedman (1979) reported a negative but relatively small *ES* (-0.24) when college text included more versus fewer grammar errors (this was the smallest effect obtained for grammar).

It is not clear why scorers would be less influenced by poor handwriting or spelling and grammar errors when scoring the writing of college students. It is possible that for handwriting, the only presentation factor where we actually

tested a grade-related difference, scorers assume that college students produce less legible writing than school-age students and are less influenced when scoring the former's writing. Such a belief is not unreasonable, as overall legibility of students' writing peaks at about fourth grade, just at the point students start to modify their handwriting to increase fluency (Graham & Weintraub, 1996). In contrast, it also seems reasonable to assume that scorers would expect more polished writing from college students (including more legible text and fewer spelling and grammar errors) and would be more likely to penalize them when this is not the case. Additional research is needed to verify differential grade effects and to explore possible reasons for such effects.

Another important area for future research is to examine the impact of handwriting, spelling, grammar, word-processing printed text, and other presentation factors on text produced by elementary grade students. Only two studies (Klein & Taub, 2005; Russell & Tao, 2004b) reviewed here were conducted with such students. Presentation effects may be even more pronounced for these students, as they are still in the process of gaining mastery over handwriting, typing, spelling, and grammar (Graham, 2006). Consequently, their papers may be less legible and include even more errors, increasing the likelihood that scorers will penalize these writers for their written flaws. Of course, this is even more likely for students with special needs (Graham & Harris, 2011).

It is also necessary to determine whether the presentation factors studied in this review interact with characteristics of the scorer or the writer. There is some limited data demonstrating such interactions do in fact exist. For example, Huck and Bounds (1972) found that scorers with messy handwriting were not influenced by text legibility when scoring students' writing, whereas scorers with good handwriting were. Other researchers found that handwriting presentation effects were mediated by knowledge of the writer. This included knowledge about the attractiveness of the writer if she was female (Bull & Stevens, 1979) and knowledge about the writer's previous achievement (Chase, 1979). This later finding clearly has important implications for students with special needs, as presentation effects may be even more pronounced for these students if the grader knows their identity. These findings need to be replicated and other mediating factors identified.

Implications for Practice

The findings of this meta-analysis raise concerns about what is assessed when the writing of students with and without special needs is scored. It appears that more than just the content or quality of the message is evaluated, as handwriting legibility, spelling or grammar errors, and computer-printed text significantly influence students' writing scores. Given the magnitude of the average weighted *ES* for these four presentation factors, an important goal in assessing students' writing is to minimize their influence as well as eliminate the influence of other factors that mediate these effects (e.g., knowledge about the writer). This is especially important with school-aged youngsters, as each study in this review conducted with text written by these students resulted in a negative effect (see Table 1). This becomes even more critical when the consequences of the assessment are high. This includes writing assessment used to determine whether students receive extra writing assistance (e.g., special education services), move from one grade to the next, graduate from high school, and attend a particular college or enter a specific trade. It also includes high-stakes assessments used to judge how well teachers, schools, states, or nations are doing in teaching students how to write.

Although the consequences are less momentous, we would also argue that presentation effects need to be minimized when writing is used as a means for assessing students' content knowledge and when grading an individual piece of writing written by students with and without disabilities, as the grades that students receive in courses extend beyond the classroom. It must further be recognized that presentation effects likely operate conjointly, working together to bias the scoring of students' writing. For example, students with poor handwriting are often poor spellers too (Graham & Harris, 2011).

Presentation effects are further complicated by the writing medium. As this review demonstrated, word-processing printed text is scored more harshly than handwritten text. Nevertheless, word processors typically contain software, such as spelling and grammar checkers, that may limit other presentation effects. Complicating the situation even further, the same writing assessment (e.g., a written report for a class) may involve some students creating a handwritten paper and others producing their paper on a word processor, creating multiple as well as contradictory biases in the scoring of the same writing assignment.

One potential means for controlling presentation effects is to blind all papers to be scored. This minimizes the chance that knowledge about the writer will interact with presentation factors to bias the assessment of students' writing. A second possible solution involves training scorers about the influence of each presentation effect. For instance, Russell and Tao (2004a) provided scorers with additional training, beyond the three hours of training they received in how to use a specific rubric to score students' writing. This additional training involved reviewing past research on the biasing effect of presentation factors (e.g., word-processing printed text, spelling errors), examining and scoring students' papers to determine how these effects influence the scorer, recommending that raters keep a mental count of the number of errors they observe while reading a paper to be scored, and encouraging scorers to think carefully about the factors that influence their

FOCUS ON EXCEPTIONAL CHILDREN

evaluation before assigning a final score. With such extra training, presentation effects were nonexistent (*ES* = -0.03 for writing quality). Similar findings were reported by Powers et al. (1994). We would, however, like to note that we think that it is unlikely that presentation effects will disappear by just providing regular training on how to score students' writing. Such training did not eliminate such effects in any of the studies reviewed here (see Table 1).

Lastly, the biasing effects due to handwriting, spelling, and grammar can be eliminated by typing papers and correcting all errors before papers are scored (researchers often do this). This is an expensive and time-consuming process. It is unlikely that individual teachers will pursue such a solution for either students with or without disabilities, especially if their students write frequently. While those who administer and score high-stakes writing assessments (e.g., college entrance exams, graduation tests, state competency exams, national tests) are also likely to resist such a solution, it does provide a means for making such assessments more valid and fair.

REFERENCES

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

Arnold, V., Legas, J., Obler, S., Pacheco, M., Russell, C., & Umbdenstock, L. (1990). *Do students get higher scores on their word-processed papers? A study of bias in scoring hand-written versus word-processed papers.* Unpuiblished manuscript, Rio Hondo College, Whittier, CA.

*Bull, R., & Stevens, J. (1979). The effects of attractiveness of writer and penmanship on essay grades. *Journal of Occupational Psychology, 52*, 53–59.

*Chase, C. (1968). The impact of some obvious variables on essay test scores. *Journal of Educational Measurement, 5*, 315–318.

*Chase, C. (1979). The impact of achievement expectations and handwriting quality on scoring essay tests. *Journal of Educational Measurement, 16*, 39–42.

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

Cutler, L., & Graham, S. (2008). Primary grade writing instruction: A national survey. *Journal of Educational Psychology, 100*, 907–919.

*Freedman, S. (1979). How characteristics of student essays influence teachers' evaluations. *Journal of Educational Psychology, 71*, 328–338.

Gewertz, C., & Robelen, E. (2010, September 15). U.S. tests awaiting big shifts: Most states part of groups winning federal grants. *Education Week, 30*(3), 1, 18–19.

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

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

Graham, S., & Harris, K. (2011). Writing. In R. Allington & A. McGill-Franzen. (Eds.), *Handbook of reading disabilities research* (pp. 232–241). Mahwah, NJ: Erlbaum.

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., & Perrin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology, 99*, 445–476.

Graham, S., & Weintraub, N. (1996). A review of handwriting research: Progress and prospect from 1980 to 1993. *Educational Psychology Review, 8*, 7–87.

Hedges, L., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

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

*Huck, S., & Bounds, W. (1972). Essay grades: An interaction between graders' handwriting clarity and the neatness of examination papers. *American Educational Research Journal, 9*, 279–283.

James, W. (1927). The effect of handwriting upon grading. *The English Journal, 16*, 180–185.

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

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

Lederer, R. (2000). *The bride of anguished English.* New York: St. Martin's Press.

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

Markham, L. (1976). Influences of handwriting quality of teacher evaluation of written work. *American Educational Research Journal, 13*, 277–283.

*Marshall, J. C. (1967). Composition errors and essay examination grades re-examined. *American Educational Research Journal, 4*, 375–385.

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

National Commission on Writing. (2004). *Writing: A ticket to work or a ticket out: A survey of business leaders.* Retrieved from http://www.collegeboard.com/prod_downloads/writingcom/writing-ticket-to-work.pdf

National Commission on Writing. (2005). *Writing: A powerful message from state government.* Retrieved from http://www.collegeboard.com/prod_downloads/writingcom/powerful-message-from-state.pdf

*Powers, D., Fowles, M., Farnum, M., & Ramsey, P. (1994). Will they think less of my handwritten essay if others word process theirs? Effects on essay scores of intermingling handwritten and word-processed essays. *Journal of Educational Measurement, 31*, 220–233.

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

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

*Scannell, D. P., & Marshall, J. C. (1966). The effect of selected composition errors on grades assigned to essay examinations. *American educational Research Journal, 3*, 125–130.

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

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

*Sweedler-Brown, C. (1991). Computers and assessment: The effect of typing versus handwriting on the holistic scoring of essays. *Research and Teaching in Developmental Education, 8*, 5–14.

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

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

*Yeh, S. S. (1998). Validation of a scheme for assessing argumentative writing of middle school students. *Assessing Writing, 5*, 123–150.