The Effect of Confounding Writing Style with Writing Content in Constructed-Response Items

Robert W. Lissitz, William D. Schafer, & Phill Gagné

Department of Measurement, Statistics and Evaluation

University of Maryland, College Park

April 24, 2003

Paper to be presented at the annual meeting of the National Council for Measurement in Education, April 22-24, 2003, Chicago.¹

¹This research was carried out by the Maryland Assessment Research Center for Education Success with funding provided through the Maryland State Department of Education. The opinions are solely those of the authors.
Most school systems and states use constructed-response items in classroom assessments and in large-scale assessments. These constructed-response items are graded partially on content and partially on writing style (punctuation, grammar, spelling, etc.) or they are graded exclusively for content. Of concern to education policy makers, to teachers, and to students is the extent to which content grades are actually based solely on content. When grading the content of a constructed response, graders do their best to evaluate the work with content as their only criterion, but research has suggested that several factors not meant to be graded can and often do influence such evaluations. Scores on writing prompts have been found to differ as a function of gender, race, and handwriting, just to name a few of the potential confounds that have been found in writing assessment.

Engelhard, Walker, Gordon, and Gabrielson (1994) examined the effects of prompt, type of knowledge, gender, and race on the grades received on essays written by 170,899 eighth graders. Narratives received higher scores than descriptive tasks which received higher scores than expository prompts. Essays that drew from direct experience were rated more highly than those for which knowledge of the world was required. These types of differences in scores, however, would potentially confound only comparisons across groups who were given different types of items, an unusual situation for K-12 assessments.

Other threats to the validity of rater scores have also been studied. Engelhard et al. (1994) found that there were significant differences in scores as a function of gender and as a function of race, with girls receiving higher scores than boys, and White students receiving higher scores than Black students. In a study of 366 high school students, Gyagenda and Engelhard (1998) found similar gender differences on holistic scores and on the subscores of content/organization, style, conventions, and sentence formation. In a meta-analysis of 165
studies, Hyde and Linn (1988) found that girls have historically performed better on verbal tasks than boys, but since 1973, the difference has declined appreciably.

Chase (1986) found significant differences in writing scores as a function of gender and race but in the form of interactions with handwriting and rater expectations. Eighty raters were given the same essay with varying handwriting quality and varying information about the writer. When handwriting was poor and expectations were low, Black females and White males received higher scores than students purported to be Black males or White females. By contrast, when expectations were high, Black males and White females received higher scores than Black females and White males.

Wolfe and Kao (1996) investigated non-content factors that affect the variance of scores assigned to essays. Thirty-six raters were asked to grade 24 essays while verbalizing their thoughts during the grading process. Raters who expressed more thoughts regarding the writing style of the paper and the writer’s ability to tell a story had significantly higher variance in the scores they assigned to the essays.

Minor errors in writing style were found not to affect the scores assigned to the GRE Writing Assessment and found not to affect application decisions when faculty on graduate admissions committees were allowed access to the essays (Powers & Fowles, 2002). The original essays were regraded after mechanical errors were introduced into the writing, resulting in a mean difference of only .06 (in favor of the regraded essays) and a correlation between the scores of .92. Furthermore, the availability of the actual essays (rather than having only the essay scores available) did not significantly contribute to explaining the variability in committee members’ admission decisions for applicants.
The present study also focused on the potential confounding of writing style in essay grades for which only content was to be graded. Of specific interest was addressing the issue of whether the content of essays in the Maryland State Performance Assessment Program (MSPAP) were being graded according to writing style, when writing style was considered construct-irrelevant. Essays which originally received low scores were altered to improve the writing style, and essays originally receiving high scores were altered to degrade the writing style. All essays were then re-graded with the same scoring rubric, and the new grades were compared to the old grades to determine whether writing style affected the grades.

Method

The essays used in the study were originally written by fifth-graders responding to items as part of MSPAP. A random sample of 130 answer booklets was drawn from those available, with each booklet containing 86 items which called for a written response of some sort. Half of these booklets had received grades of “Unsatisfactory” in all contents and half had originally received a grade of “Satisfactory” in all contents. Each year MSPAP uses an equating cluster, a version of the assessment identical to a version from the previous year. The booklets used in the present study were from the equating cluster, a design feature which facilitated keeping the raters blind to the study.

Several retired principals selected by the Maryland State Department of Education (MSDE) were trained by one of the authors to alter writing style without altering content, with the magnitude of suggested alterations ranging from changing the spelling of some words to reordering entire paragraphs. For the 65 booklets originally rated as “Satisfactory,” specific changes to the text were applied to degrade the writing style for all of the responses. For the 65 “Unsatisfactory” booklets, specific changes to the text were applied to improve the writing
throughout each booklet. The specific methods used to improve or degrade the booklets were based on a literature search and are listed in the Appendix. The booklets were then given to another group of people who copied the answers into new booklets, making all of the suggested alterations to the writing style, resulting in 65 poorly written booklets with comparatively solid content and 65 well written booklets with comparatively meager content.

The raters were Maryland teachers employed and supervised by a professional scoring service hired by MSDE. Following standard MSPAP procedures, the papers were rated in random order by four teams, one scoring writing and language usage, one scoring mathematics, one scoring science, and one scoring social studies. Scoring of reading was divided among the other teams. Each team consisted of several tables and a leader; check sets and read-behinds were routinely scored and the performance of raters monitored. Further details of the MSPAP scoring process are available at the mdk12.org website.

Only the primary supervisor at the scoring site was aware of the study; all scorers and table leaders were blind to the research. The test answer booklets with the original responses were interspersed and graded along with all of the other booklets for the MSPAP in the year 2000; these ratings were those that were used in booklet selection. The altered booklets were graded by a new set of raters, along with all of the other booklets for the MSPAP in the year 2001.

Results

With the exception of five of the items, each item contributed uniquely to one of six area scores: writing (3 items), language usage (8 items), reading (7 items), social studies (17 items), mathematics (30 items), and science (16 items). Items involved in more than one content area were deleted from the analysis.
Meta-analysis was used to study the remaining 81 items, with the item as the unit of analysis. For each item, an effect size (d) was calculated as the mean across examinees of the scores from the “good writing” condition minus the mean of the scores from the “poor writing” condition divided by the pooled standard deviation for both conditions. Listwise deletion was used to deal with missing data, which led to the number of examinees varying from 96 to 130 across the 81 items.

Each effect size was weighted by the inverse of its variance. The average effect size was .066 ($\chi^2(1) = 125.28, p < .01$). There was, however, significant variation of effect sizes about the average ($\chi^2(80) = 682.54, p < .01$). The effect sizes were therefore partitioned using the six content areas. The reduction in effect size variation was significant ($\chi^2(5) = 510.15, p < .01$). Although there remained significant variation of effect sizes around the six means ($\chi^2(75) = 172.39, p < .01$), further partitioning was not attempted because homogeneity was achieved in four of the six areas.

Table 1 provides number of items, effect sizes, and standard errors for all six areas. All effect sizes were tested with a chi-square analysis and found to be significant at the .01-level. The effect sizes for language usage (d = .805, S.E. = .036) and writing (d = .472, S.E. = .053) were large and medium (Cohen, 1988), respectively, but the effect sizes for all of the other four content areas were less than half of Cohen’s “small” effect size of .2.

<table>
<thead>
<tr>
<th>Scoring Area</th>
<th>No. of Items</th>
<th>Ave. Effect Size</th>
<th>Effect S. E.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language Usage</td>
<td>8</td>
<td>0.805</td>
<td>0.036</td>
</tr>
<tr>
<td>Writing</td>
<td>3</td>
<td>0.472</td>
<td>0.053</td>
</tr>
<tr>
<td>Reading</td>
<td>7</td>
<td>0.099</td>
<td>0.03</td>
</tr>
<tr>
<td>Social Studies</td>
<td>17</td>
<td>0.076</td>
<td>0.017</td>
</tr>
<tr>
<td>Science</td>
<td>16</td>
<td>0.051</td>
<td>0.013</td>
</tr>
<tr>
<td>Mathematics</td>
<td>30</td>
<td>0.025</td>
<td>0.008</td>
</tr>
</tbody>
</table>
Discussion

The magnitude of the effect sizes for language usage and writing are not surprising. These are areas in which writing style is intended to affect grades, so it is consistent with expectations that these effect sizes were appreciable. For the other four content areas (reading, social studies, science, and math), the effect sizes were less than “small.” Grades for these content areas should be unaffected by writing style, and these results indicate that the graders were successful at grading the content of the responses without appreciable consideration for writing style. This study provides considerable support for the belief that content scores are only marginally affected by style considerations, when scoring is done by well trained professionals.
References


Appendix

A Quick Guide to Improving/Worsening Writing Style

I. Punctuation, Spelling, and Grammar

1) Commas: Adding commas where needed is an easy way to improve the flow of the writing. Dropping commas is likewise an easy way to diminish the quality of the writing. It is unusual, however, for comma mistakes to be excessive overuse. Misuse, such as with conjunctions and appositives, and underuse are more typical errors; in fact, it would not be unreasonable to take all of the commas out of one or two responses.

2) Periods: Nearly all, if not all, of the students know how to use periods, and most are so uncertain about the use of semicolons and colons that they would not use them in a graded writing situation. Adding a period where one is needed can greatly improve writing quality, but it can also make the flow choppy, so check the appropriateness of a period against the use of a conjunction. It is okay to omit a period occasionally, but if you do, omit one from a sentence that does not end a paragraph. A missing period at the end of a paragraph tends to be attributed either to a mental lapse or to excessive speed rather than to a lack of writing skill.

3) Spelling: When worsening writing, make reasonable choices of words to misspell. No one is going to misspell “the”, and long but phonetic words such as “establishment” are also not likely to be misspelled. Spelling errors involving the suffix “tion” typically occur en masse for a given person, so it is unlikely for “instruction” to be misspelled at the end but not “institution”. Less complicated words to misspell feature potential “ie” vs “ei” confusion and “s” vs “ss” confusion.

4) Verb tense: Correcting inaccurate verb tense is a quick way to improve writing quality. Changing verbs to an incorrect tense is usually an obvious error, but sometimes it only serves to alter the meaning of a sentence. For example, changing “I went to the store” to “I will go to the store” has no effect on the accuracy of the sentence unless, up to that point and for some time thereafter, I give an account of my day in the past tense.

5) Prepositions: Add, change, or drop prepositions and conjunctions on which the content does not depend but on which the clarity or grammatical accuracy does. For example, dropping the conjunction in a compound sentence and breaking it into two simpler sentences can sometimes add confusion and oversimplicity to a previously clear and sophisticated sentence, without affecting the content. Changing “She ran to the race” to “She ran in the race”, however, alters the meaning of the sentence, so avoid trying to worsen the writing by making such changes.

6) Complete sentences: Some responses to prompts that are not formally testing writing include only content words, so these could be improved by filling in appropriate words to make a complete sentence or two. For responses written as complete sentences, you can decrease the writing quality by removing all of the grammar and structural components, but be careful not to affect the content with such alterations.

II. Word Choice

1) Complexity: Replacing sophisticated words and phrases with simpler or less vivid synonyms has a negative effect on writing quality. “Contrary to” becomes “different than”, “ecstatic” becomes “happy”, “innumerable” becomes “a lot”, etc. To improve writing, apply the reverse alteration.
2) **Variety**: When the same adjective is used repeatedly in a paper, better writers will use a few synonyms instead of writing the same word repeatedly. For example, where a good writer would use “clean”, “spotless”, “immaculate”, and “dirt-free”, a bad writer might just use “clean” four times in a row.

3) ** Appropriateness**: Variety or a momentary lapse in thought can sometimes lead to the use of a word that does not work as well, or even at all, relative to a synonym of that word. For example, it would be a mistake for “I was soaking wet when I came in from the heavy downpour” to be written as “I was moist when I came in from the heavy downpour.”

4) **Homophones**: Writing can be made worse by replacing a word with a correctly spelled, but incorrectly used, homophone or near homophone. For example, “According to old lore, that house is haunted” could be changed to “According to old lure, that house is haunted.” As young writers are likely to have made this mistake on their own, it might prove more useful to look for opportunities to improve the response by using the correct version of a homophone.

III. Organization

1) **Order of paragraphs**: Switching the order of non-introductory and non-concluding paragraphs can disrupt the flow of a piece of writing, but avoid making such changes for writing tasks that have an obvious chronology. For example, even a bad writer would not organize an account of American history to describe the Civil War before the Revolutionary War.

2) **Order of sentences**: Sentences within paragraphs and between paragraphs can be switched so long as the content is not manipulated. Some sentence switches might have too subtle an effect for our purposes, so be sure to compound subtle switches. Another recommended procedure for reordering sentences is to pull a sentence from one paragraph and replace it in a different paragraph. To improve writing, look for sentences that would make better sense in a different order or even in a different paragraph. When attempting to improve by moving a sentence from one paragraph, however, be careful not to counter the improved organization of thought with diminished flow of the writing.