Research Memorandum

"Wordiness": A Selective Review of Its Influence, and Suggestions for Investigating Its Relevance in Tests Requiring Extended Written Responses

Donald E. Powers

Research & Development

February 2005
RM-04-08

"Wordiness": A Selective Review of Its Influence, and Suggestions for Investigating Its Relevance in Tests Requiring Extended Written Responses

Donald E. Powers
ETS, Princeton, NJ

February 2005

As part of its educational and social mission and in fulfilling the organization's nonprofit charter and bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

www.ets.org/research/contact.html

Abstract

This paper provides a brief review of the relationship between response length and response quality for constructed-response measures, primarily essay responses for tests of writing skill. Some suggestions are made for study designs that may be useful for studying the influence of response length as a threat to the validity of constructed-response measures.

Key words: Test validity, constructed responses, essay tests, response length, wordiness

For a variety of reasons, assessments that require examinees to perform or to produce/construct responses are becoming increasingly popular (Aschbacher, 1991). They are popular in part because, when compared with traditional multiple-choice measures, they are typically regarded as being *direct*. That is, the skills and abilities of interest are often readily apparent in the resulting products or performances. (I am speaking here primarily of written, natural language responses that require test takers to explain a phenomenon, discuss an issue, or develop an argument; for instance, the responses are not those that entail quantitative answers to mathematical problems such as proofs.)

Some forms of performance assessments and constructed-response tests have also been referred to as *free*-response measures and with good reason. Whereas traditional multiple-choice tests permit only a selection from several prespecified options, constructed-response measures typically allow examinees considerably more latitude with regard to both what they say and how they say it. Clearly then, one degree of freedom for many constructed-response measures is the sheer amount that test takers can offer in response to a question or prompt: Some test takers may opt for (or be able to produce only) relatively short replies; others may prefer (and be capable of constructing) answers that are considerably longer. Thus, unless examinees are directed (or constrained) to produce responses of a specified size, response length is likely to vary significantly among examinees. As this paper shows later, judgments of response quality are often highly related to response length. Our interest here is the degree to which this variation is construct-relevant versus the extent to which it may, as the title of this paper suggests, simply constitute extraneous "wordiness."

Response Length and Perceived Quality

Irrespective of the cause, there is ample evidence of a strong relationship between the length of examinee-constructed responses and the evaluations they receive. In fact, Rafoth and Rubin (1984) asserted, albeit on the basis of fairly limited research (Page, 1968; Stewart & Grobe, 1979), that "…composition length is well established as the single most powerful predictor of composition quality ratings" (p. 447).

This strong relationship is sometimes interpreted as being irrelevant and unwanted, or even worse, biased in favor of longer essays. Some research seems to support this view. Charney (1984), for instance, suggested that essay readers may focus on characteristics (spelling errors and essay length, for example) that, while easy to notice, are largely extraneous to the ability to

write well. Consistent with this view is an unsubstantiated notion that some large-scale writing assessment programs encourage such rapid evaluation of responses that readers cannot render informed judgments and instead resort to scoring that is based primarily on such readily noticeable features as essay length.

On the other hand, scoring guidelines often require examinees to elaborate the views they espouse and to support their opinions with well-reasoned examples. The strong relationship between response length and response quality may, therefore, be entirely appropriate. Even the briefest perusal of constructed-response scoring guides suggests that length of response is certainly not entirely irrelevant. The scoring guides used for National Assessment of Educational Progress (NAEP) mention the desirability of elaborated responses that go beyond the essentials, as well as the insufficiency of overly abbreviated responses. In another context, essays written for the Praxis Series™ certification tests for beginning teachers are downgraded if they do not illustrate key ideas adequately or if they furnish little relevant detail. Similarly, the Graduate Record Examinations® (GRE®) analytical writing assessment also stresses the need to buttress assertions with compelling reasons, examples, and supporting detail. As Breland, Bonner, and Kubota (1995) opined, the relationship between essay length and score is mainly a natural side effect of the need to provide supporting detail:

> Using supporting materials, for example, requires that words be written to explain the relevance of such materials. And it is impossible to fully develop an essay without writing a sufficient number of words. [Thus] extremely brief essays are probably very unlikely to receive high holistic scores. (p. 14)

In addition to scoring guidelines, the way in which essay readers are typically trained to deal with response length is also revealing. For example, in large-scale writing assessments, like those that are part of the Graduate Management Admission Test® (GMAT®) and the GRE General Test, prospective readers are instructed that, because some short essays deserve high scores and some long ones low scores, they should *not* judge essays by their length.

Finally, the (unofficial) advice that test preparation firms give to those preparing for writing tests like the GRE analytical writing assessment is also enlightening. In one test preparation publication, prospective GRE examinees are told that:

> An essay that is concise and to the point can be more effective than a long-winded, rambling one. On the other hand, a longer essay that is nevertheless articulate and that includes many insightful ideas that are well supported by examples will score higher than a brief essay that lacks substance. (Stewart, 2000, p. 6)

Elsewhere in another non-GRE program publication, test takers are advised simply to make certain that they "…fill up enough space" (Lurie, 2000, p. 232).

Students preparing to take the SAT II: Writing Test are told by one test preparer that "…economy is a virtue," and although "…quantity is far less important than quality," a single paragraph is probably not sufficient to develop ideas adequately (Ehrenhaft, 1998, p. 60). Another publication's advice is similar: "One or two paragraphs, no matter how sophisticated, just won't cut it…," but neither will a long essay that's "filled with fluff" (Kaplan, Inc., 2001, p. 29).

Thus far, then, it appears that quantity cannot be readily construed as being either entirely relevant or entirely irrelevant to the construct of writing proficiency. Moreover, even if irrelevant, there would still be some uncertainty about the import of this source of invalidity. It is one thing if essay length is primarily an indication of verbal fluency (instead of writing skill). It is another, however, if it mainly represents a tendency to include extraneous "padding." To quote a 1950's advertising jingle that touted the length of one manufacturer's king-size cigarette, "It's not how long they make it, it's how they make it long." So it is (or at least *should be*) with constructed responses.

The objective of this monograph is to suggest a strategy for discounting the threat that response length poses to the evaluation/interpretation of constructed responses. That is, the aim is to rule out, or at least render implausible, the likelihood that constructed response evaluations reflect, primarily, the length of responses and not their quality, however quality may be defined. The discussion here is restricted primarily to essay responses that are intended to reflect writing skill. I believe, however, that the methods discussed below could be applicable to other kinds of constructed responses (see Table 1).

Table 1

Summary of Selected Studies Addressing Relation Between Length and Quality of Constructed Response

Study	Kind of constructed response	Type of scoring	Sample	Selected findings
Breland, Bonner, & Kubota (1995)	Impromptu essays written for the College Board English Composition Test	Holistic. Additional holistic and analytical ratings were also obtained.	400 secondary school students	Number of words written correlated .72 with essay scores. In regression analyses, the number of words written was the strongest predictor of scores, even when other characteristics of essays and SAT verbal scores were included as predictors. Words written accounted for 27% of the variance in essay scores; next most predictive variable accounted for 11%.
Breland, Danos, Kahn, Kubota, & Sudlow (1991)	Two College Board Advanced Placement Program® history examinations (free-response portions)	Holistic. Special historical content scores were based on number of main points and amount of supporting evidence; "English scores" were based on reading the historical essay responses as if they had been written for English courses.	Approximately 800 secondary school students taking one of two AP history exams	The correlation between number of words written and free-response scores was .57 for U.S. history and .61 for European history. The corresponding correlations between words written and historical content scores were .50 and .63. Correlations between words written and English scores were .33 and .43 for two parts of the U.S. history exam and .58 and .66 for European history.
Breland & Jones (1982)	Essays written for the College Board English Composition Test	Holistic	806 secondary school students	Essay length (number of lines written) correlated .51 and .50 with two different holistic scores.
Frase, Faletti, Ginther, & Grant (1999)	Expository essays written for the TOEFL Test of Written English™ (TWE®)	Holistic	Approximately 1,700 essays written by international students from 5 language groups	The number of words and the average length of words taken together correlated greater than .80 with essay scores. The number of words made a greater contribution to prediction than did word length, with partial correlations ranging from .65 to .82 in the samples.

Table 1 (continued)

Study	Kind of constructed response	Type of scoring	Sample	Selected findings
Kaplan, Wolff, Burstein, Lu, Rock, & Kaplan (1998)	Expository essays written for the Praxis Series: Professional Assessments for Beginning Teachers®	Holistic	1,014 essays written by undergraduate teacher trainees	The correlation of essay scores with number of words was .68; with the fourth root of the number of words, .70. (As essay length exceeded 400 words, the relationship diminished.)
Norton (1990)	Essay written outside class as a course assignment. A limit of 1,250 words was imposed.	Graded by course tutors according to unspecified criteria	98 first-year undergraduate psychology students	The correlation between number of words and grade assigned to the essay was .13.
Page (1994)	Narrative essays written for the National Assessment of Educational Progress (NAEP)	Holistic	599 high school seniors	When scores were predicted using 30 some features of essays, the fourth root of the number of words in the essay had a standardized regression weight of .68.
Powers, Fowles, Farnum, & Ramsey (1994)	Expository essays written for the Praxis Series: Professional Assessments for Beginning Teachers	Holistic	128 essays written by 32 undergraduate teacher trainees, 1 by computer and 1 by hand and each subsequently converted to the other mode by the researchers	Word count correlated "in the .60s" regardless of the mode in which essays were produced or scored.
Powers, Fowles, & Welsh (1999, 2001)	Papers written for undergraduate course assignments	Holistic	Approximately 1,800 upper-level undergraduate students	The correlation between number of words in a paper and the holistic evaluation it received was .51 for each of two papers submitted by study participants. (One was typical of their work and the other was of lower quality.)
Russell & Haney (1997)	A performance writing assessment requiring an extended written response to a question about an artist's mural	Holistic	120 middle school students	Length of responses, both in terms of number of words and length of words, correlated .63 with essay scores. Students writing on computer wrote almost twice as much as those who handwrote.

Relevant Research

Some of the evidence relating quantity and quality of constructed responses is indirect. For example, when test takers are given more time, they write more and their scores are higher also (Powers & Fowles, 1996). Also, when skilled word processors write, they produce longer essays (and get higher scores) when composing on computer than when handwriting (Wolfe, Bolton, Feltovich, & Niday, 1996).

Table 1 summarizes some of the research that has revealed a relationship between the length of constructed responses and the evaluations they receive. Much of the research has focused on tests of writing skill, but some has included achievement tests. The independent variable of interest—quantity—has been operationalized in slightly different ways in the studies. Sometimes exact word counts are used or the number of lines of text counted. Other times, slightly cruder measures are employed such as estimates of total words, as computed from number of lines and average words per line. Some studies have also incorporated *word length* in the calculations.

A majority of the studies have investigated holistic or impressionistic ratings based on various scoring guidelines. Some studies have, in addition, examined specially devised scores reflecting essay content, for example. Most of the participants in the cited studies have been either secondary school students or college undergraduates. In most cases, the constructed responses were obtained from large-scale operational testing programs, but in a few instances, the responses were based on classroom assignments.

Over all of the studies reviewed, the correlations of response length with response quality range from a low of 0.13 to a high of 0.80. Correlations in the 0.50s to the 0.70s are common, suggesting that response length alone can very often account for up to half the variation in constructed-response scores. (Of course, response length may be a proxy for other more relevant factors.) It is interesting to note that the lowest correlation (0.13)—a clear outlier—was noted in a study in which writers were constrained (or at least encouraged) to keep their responses to a course assignment to a minimum number of words.

Most of the studies reported only linear relations between response length and response quality. A few, however, explored nonlinear relations, finding that beyond a certain number of words, the relationship between length and quality seemed to diminish somewhat.

Several studies investigated the role of a variety of other response characteristics, in addition to response length. Most of these examined the characteristics in combination, finding usually that response length was a very strong (and often the best) predictor of test scores, even when considered jointly with other essay features.

Possible Approaches

A variety of approaches could be used to establish the relevance of response length in the evaluation of constructed responses. Both correlational and experimental approaches could be informative, the former for *ruling out* response length (and various other factors) as causes of response quality (by virtue of their lack of relationship) and the latter for establishing more definitive causal links. Some of these factors may be entirely construct-relevant, some only partially relevant, and some largely irrelevant and unwanted. For responses intended to reflect writing skills, construct-relevant variables might include:

- knowledge of the conventions of standard written English
- the ability to think critically
- the ability to articulate complex ideas clearly and effectively (measured, for example, orally, instead of in writing)
- the ability to generate reasons, hypotheses, or examples
- the extent to which the peers and teachers regard the writer as being skilled
- independent judgments or measures of writing skill (e.g., editing) that do not involve response length

Variables that might be considered partially relevant to a construct of writing might be:

- knowledge of the topic
- verbal fluency or aptitude
- creativity

Examples of variables that would probably be considered to be largely construct irrelevant would be:

- interest in the topic

- the ability to elaborate without having any significant knowledge of a topic
- word processing (or handwriting) skills

Response length could be considered in conjunction with these variables in order to determine the factors for which response length may be a proxy. As stated above, the main value of such correlational studies would be to rule out some variables as underlying causes, by virtue of their lack of relationship with response length or with response quality.

While correlational studies would be suggestive, experimental studies could, by controlling other factors, help to establish any causal link between response length and essay scores. Such studies might involve the manipulation of the following variables:

- instructions to test takers (e.g., by encouraging them to write either succinct or fully developed responses)
- time limits or physical constraints (e.g., by controlling the amount of time allotted or space available for writing)
- the length of responses produced by examinees (e.g., by inserting or deleting material)
- instructions to readers/scorers (e.g, to ignore or to discount response length)

Some of these manipulations would entail multiple tactics. For example, response length could be varied in several ways. Responses could be artificially shortened by simply deleting blocks of text, by deleting complete sentences throughout the response, or by deleting words within each sentence within a response. Responses could be lengthened by rephrasing sentences or by introducing redundancies. Each tactic might have distinct consequences and would presumably require different interpretations. In any event, these manipulations would need to proceed very carefully so as not to inadvertently improve or impair the overall quality of the writing.

To argue successfully that essay scores reflect mainly response length, it seems that the following conditions should be expected to hold:

- when examinees whose responses get *low* scores are allowed or encouraged to write longer responses, their scores should *increase* appreciably

- when those whose responses get *high* scores are restricted or encouraged to write short responses, their scores should *decrease* appreciably
- the variation among scores should *decrease* significantly when response length is controlled or held constant
- the variation among scores should *increase* significantly when response length is completely uncontrolled
- the artificial lengthening or shortening of responses (in a way that does not unduly affect other aspects of response quality) should have an appreciable effect on scores.

To the extent that these conditions fail to hold, then response length would be rendered less plausible as a threat to the desired interpretation of test scores as reflections of writing skill.

An Illustration

The GRE Web site (http://www.GRE.org) contains six GRE test taker essays that exemplify each of the six possible scores on the GRE analytical writing assessment. The prompt on which test takers were asked to "present your perspective, using relevant reasons and/or examples to support your views" was: "In our time, specialists of all kinds are highly overrated. We need more generalists—people who can provide broad perspectives."

The essays vary considerably in length, ranging from 108 to 623 words. Scores are perfectly rank-ordered according to the number of words in the essays. The author decreased the length of each of these essays by taking only the first portion of each one so as to create responses having approximately equal numbers of words, ranging from 108 to 128 (see Appendix A). Although it now seems much more difficult to differentiate the quality of these responses, it still seems possible to discriminate at least the best writers from the worst. The reader is encouraged to rank order each of the shortened responses and to check his/her rankings against the scores that were awarded to the original, longer essays (see Appendix B for the rating key.)

References

Aschbacher, P. R. (1991). Performance assessment: State activity, interest, and concerns. *Applied Measurement in Education, 4,* 275-288.

Breland, H. M., Bonner, M. W., & Kubota, M. Y. (1995). *Factors in performance on brief, impromptu essay examinations* (College Board Rep. No. 95-04; ETS RR-95-41). New York: College Entrance Examination Board.

Breland, H. M., Danos, D. O., Kahn, H. D., Kubota, M. Y., & Sudlow, M. W. (1991). *A study of gender and performance on Advanced Placement history examinations* (College Board Rep. No. 91-04; ETS RR-91-61). New York: College Entrance Examination Board.

Breland, H. M., & Jones, R. J. (1982). *Perceptions of writing skills* (College Board Rep. No. 82-4; ETS RR-82-47). New York: College Entrance Examination Board.

Burstein, J., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. (1998, April). *Computer analysis of essays.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the teaching of English, 18,* 65-81.

Ehrenhaft, G. (1998). *Barron's how to prepare for the SAT II writing* (2nd ed.). New York: Barron's.

Frase, L. T., Faletti, J., Ginther, A., & Grant, L. (1999). *Computer analysis of the TOEFL Test of Written English (TWE)* (TOEFL Research Rep. No. 64; ETS RR-98-42). Princeton, NJ: ETS.

Kaplan, Inc. (2001). *SAT II: writing* (6th ed.). New York: Simon & Schuster.

Kaplan, R. M., Wolff, S. E., Burstein, J. C., Lu, C., Rock, D. A., & Kaplan, B. A. (1998). *Scoring essays automatically using surface features* (GRE Board Professional Rep. No. 94-21P; ETS RR-98-39). Princeton, NJ: ETS.

Lurie, K. (2000). *Cracking the GRE.* New York: Random House.

Norton, L. S. (1990). Essay-writing: What really counts? *Higher Education, 20,* 411-442.

Page, E. B. (1968). Analyzing student essays by computer. *International Review of Education, 14*, 210-225.

Page, E. B. (1994). Computer grading of student prose, using modern concepts and software. *Journal of Experimental Education, 62,* 127-142.

Powers, D. E., & Fowles, M. E. (1996). Effects of applying different time limits to a proposed GRE writing test. *Journal of Educational Measurement, 33,* 433-452.

Powers, D. E., Fowles, M. E., 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.

Powers, D. E., Fowles, M. E., & Welsh, C. K. (1999). *Further validation of a writing assessment for graduate admissions* (GRE Board Res. Rep. No. 96-13R; ETS RR-99-18). Princeton, NJ: ETS.

Powers, D. E., Fowles, M. E., & Welsh, C. K. (2001). Relating performance on a standardized writing assessment to performance on selected academic writing activities. *Educational Assessment, 7,* 277-303.

Rafoth, B. A., & Rubin, D. L. (1984). The impact of content and mechanics on judgments of writing quality. *Written Communication, 1*(4), 446-458.

Russell, M., & Haney, W. (1997). Testing writing on computers: An experiment comparing student performance on tests conducted via computer and via paper-and-pencil. *Education Policy Analysis Archives, 5*(3). Retrieved December 16, 2004, from http://epaa.asu.edu/epaa/v5n3.html

Stewart, M. A. (2000). *GRE answers to real essay questions*. New York: IDG Books Worldwide.

Stewart, M., & Grobe, C. H. (1979). Syntactic maturity, mechanics of writing, and teachers' quality ratings. *Research in the teaching of English, 13,* 207-215.

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

Appendix A

Shortened GRE Essays

Essay Response – A

In this era of rapid social and technological change leading to increasing life complexity and psychological displacement, both positive and negative effects among persons in Western society call for a balance in which there are both specialists and generalists.

Specialists are necessary in order to allow society as a whole to properly and usefully assimilate the masses of new information and knowledge that have come out of research and have been widely disseminated through mass global media. As the head of Pharmacology at my university once said (and I paraphrase): "I can only research what I do because there are so many who have come before me to whom I can turn."

Essay Response – B

Specialists are just what their name says: people who specialize in one part of a very general scheme of things. A person can't know everything there is to know about everything. This is why specialists are helpful. You can take one general concept and divide up three ways and have three fully developed different concepts instead of one general concept that no one really knows about. Isn't it better to really know something well, than to know everything half-way.

Take a special ed teacher compared to a general ed teacher. The general ed teacher knows how to deal with most students. She knows how to teach a subject to a student that is on a normal level.

Essay Response – C

In the situation of health I feel that specialists are very important. For example if a person has heart problems, choose a heart specialist over a genral medicine Dr. However if a person is having a wide range of syptoms, perhaps choose a Dr. with a wide range of experience might be more helpful.

It also depends on the type of problem you are having. For example I would not suggest taking a troubled child to a theorpist who specializes in marriage problems. In some cases have a specialists helps to insure that you are getting the best possibly treatment. On the other hand

dealing with a person who has a wide range of experience may be able to find different ways of dealing with a particular problem.

Essay Response – D

I disagree with the statement about specialists, we need specialists who take individual areas and specialize. A generalists can pinpoint a problem. He or she cannot determine the magnitude of the problem. A specialist can find the root of the problem. When he or she has years working in that specific field. For example, when i got sick i went to a doctor. He did blood work, x-ray, talk to me, ect. He prescribed me a medicine. I got worst. So i decided to go another doctor. Now, i am doing great. A specialist knows the facts right away. Otherwise, it will take longer or not at all.

Essay Response – E

To quote the saying, "Jack of all trades, master of none," would be my position on the statement. I feel specialists in all areas of knowledge lead to a higher standard of living for everyone. Specializing in different areas allows us to use each others talents to the highest level and maximize potential. As an example, if a person required brain surgery, would they rather have a brain surgeon or a general practitioner doing the work? Clearly a specialist would do the better job and give the patient a chance at a better life.

A university education starts by laying the groundwork for general knowledge but then narrows down to a specific field.

Essay Response – F

Specialists are not overrated today. More generalists may be needed, but not to overshadow the specialists. Generalists can provide a great deal of information on many topics of interest with a broad range of ideas. People who look at the overall view of things can help with some of the large problems our society faces today. But specialists are necessary to gain a better understanding of more in depth methods to solve problems or fixing things.

One good example of why specialists are not overrated is in the medical field. Doctors are necessary for people to live healthy lives. When a person is sick, he may go to a general practitioner to find out the cause of his problems.

Appendix B

Essay Scoring Key

Essay	*Score*
A	6
B	4
C	2
D	1
E	3
F	5