Research Report

A Developmental Writing Scale

Yigal Attali

Don Powers

April 2008

ETS RR-08-19

A Developmental Writing Scale

Yigal Attali and Don Powers
ETS, Princeton, NJ

April 2008

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:

http://www.ets.org/research/contact.html

Abstract

This report describes the development of grade norms for timed-writing performance in two modes of writing: persuasive and descriptive. These norms are based on objective and automatically computed measures of writing quality in grammar, usage, mechanics, style, vocabulary, organization, and development. These measures are also used in the automated essay scoring system *e-rater*® V.2. Norms were developed through a large-scale data collection effort that involved a national sample of 170 schools, more than 500 classes from 4th, 6th, 8th, 10th, and 12th grades and more than 12,000 students. Personal and school background information was also collected. These students wrote (in 30-minute sessions) up to 4 essays (2 in each mode of writing) on topics selected from a pool of 20 topics.

The data allowed us to explore a range of questions about the development and nature of writing proficiency. Specifically, this paper provides a description of the trajectory of development in writing performance from 4th grade to 12th grade. The validity of a single developmental writing scale is examined through a human scoring experiment and a longitudinal study. The validity of the single scale is further explored through a factor analysis (exploratory and confirmatory) of the internal structure of writing performance and changes in this structure from 4th grade to 12th grade. The paper also explores important factors affecting performance, including prompt difficulty, writing mode, and student background (gender, ethnicity, and English language background).

Key words: Writing proficiency, Automated essay scoring, *e-rater*

Table of Contents

List of Tables

Page

List of Figures

Introduction

For multiple-choice tests, sophisticated methods are used to equate and maintain score comparability across test forms. The task of maintaining comparability is invariably much more difficult for performance assessments, such as writing measures, which typically rely on only a very few tasks—a situation that often precludes the use of sophisticated equating designs. A further complication is that the resulting performances on these tasks are scored subjectively by human raters, who, if not exquisitely trained, may assign ratings that drift over time or that may not be strictly comparable across tasks or across samples.

An even more daunting task in educational measurement is the maintenance of comparability over grade levels. Devising assessments to measure change and determine developmental growth has always posed significant challenges (both technical and practical) for education researchers and policy makers. Various approaches are beset with a variety of technical and logistical problems, and many are based on assumptions that are not entirely plausible. As a result, some methods may identify patterns of growth that are not realistic educationally (Petersen, Kolen, & Hoover, 1989).

The project reported here is an attempt to develop meaningfully comparable scores of essay writing performance across grade levels in a way that will enable assessment of developmental growth. Objective scoring was the key to sidestep the measurement and comparability problems regularly encountered in scoring writing assessments. The approach was to extract important features of writing via computer by using natural language processing methods. These features allow the development of objective and analytic scoring for writing performance. Furthermore, they allow the development of a unified scoring standard across topics and grade levels. The hope is that this approach will yield estimates of growth in writing skills that are judged to be educationally meaningful.

Essay Writing Assessments

Writing is a fundamental skill and an important part of school curricula from early elementary grades through high school and post-secondary education. The theoretical domain of knowledge and skills, or construct, associated with writing ability is enormously complex. Writing addresses a variety of purposes, like telling a story, informing the reader, or persuading the reader. These purposes are influenced by the audience of the writer. A variety of cognitive

skills and knowledge structures are also involved in the process of writing. Therefore, it is no surprise that writing tests do not assess the full range of the construct.

The two most popular types of writing assessment tasks are multiple-choice questions and essay writing tasks. Multiple-choice tasks typically assess the ability to improve or correct presented text. Essay writing tasks assess the ability of examinees to reflect on a given topic and then articulate their thoughts. As direct measures of writing, essay writing tasks have the potential to assess the writing construct more fully than do multiple-choice writing measures. However, in practice, the predictive validity of a multiple choice test is at least as high as a single essay test, and the reliability of the multiple-choice test is higher than a single essay test (Breland, Bridgeman, & Fowles, 1999).

One of the main complicating factors in essay tests is the complexity and cost involved in scoring. Scoring essays requires the development of scoring guides that describe for each score level (typically six) the characteristics of writing that are typical of the level. For example, the National Assessment of Educational Progress (NAEP) fourth grade narrative scoring guide (Persky, Daane, & Jin, 2003, p. 86) specifies four characteristics of an "excellent" (highest score level):

1. Tells a well-developed story with relevant descriptive details across the response.
2. Presents well-connected events that tie the story together with transitions across the response.
3. Sustains varied sentence structure and exhibits specific word choices.
4. Exhibits control over sentence boundaries; errors in grammar, spelling, and mechanics do not interfere with understanding.

However, although necessary, these guides are not sufficient for scoring individual essays. They must be supplemented by example essays at each score level in order for raters to get a sense of how to interpret the guides. In typical applications, raters are gathered for training sessions in which the use of the scoring guides is explained and exemplified. During actual scoring sessions various procedures are employed to ensure that raters' scores do not drift and keep in reasonable agreement with other raters' scores. Even with these procedures in place, the alpha reliability estimates for a two-essay assessment (each taking 30-45 minutes) with two readers per essay is only about .70 (Breland et al., 1999).

Automated Essay Scoring

Despite these difficulties in scoring essays, essay tests remain an important task in writing assessments because of their authenticity (Linn, Baker, & Dunbar, 1991) and the desire for tests that drive instruction in appropriate ways (Resnick & Resnick, 1990). In the case of writing assessments, the argument is that multiple-choice tests have the unintended side effect of focusing instruction on sentence-level problems of mechanics at the expense of the more global aspects of writing.

These difficulties in scoring have also led to a growing interest in the application of automated natural language processing techniques for the development of automated essay scoring (AES) as an alternative to human scoring of essays. As early as 1966, Page developed an AES system and showed that an automated rater is virtually indistinguishable from human raters. In the 1990s, additional systems were developed; the most prominent ones are the Intelligent Essay Assessor (Landauer, Foltz, & Laham, 1998), Intellimetric (Elliot, 2001), a new version of the Project Essay Grade (Page, 1994), and *e-rater*® (Burstein, Kuhich, Wolff, Lu, & Chodorow, 1998).

This study uses the technology of *e-rater* V.2 (Attali & Burstein, 2006). *E-rater* V.2 differs from the previous version of *e-rater* and from other AES systems in several important ways that contribute to its validity. The set of essay features used for scoring is relatively small, and all of the features are indicators of generally acknowledged dimensions of good writing (although by no means cover all aspects of good writing). Consequently, the same features are used in different scoring applications. In addition, the procedures for combining the features to obtain an essay score are simple and can be based on expert judgment. Finally, scoring procedures can be successfully applied to data from several essay prompts of the same assessment. This means that a single scoring model is developed for a writing assessment, consistent with the human rubric that is usually the same for all assessment prompts in the same mode of writing.

The feature set used in this study is based on the features in *e-rater* V.2 (see Table 1), which includes measures of grammar, usage, mechanics, style, essay length, vocabulary, and word length (see also Attali & Burstein, 2006). Essay length was used instead of the organization and development features of *e-rater* V.2 because of the very high, combined multiple correlation of these two features with essay length. In addition, it is possible in *e-rater* V.2 to use prompt-specific vocabulary usage measures. However, the computation of these measures is specific to

each prompt and therefore is not suitable for this study, which seeks to develop essay scoring standards that are uniform across prompts.

In order to compute an essay score, the feature values should be combined in some way. In *e-rater* V.2, this task is accomplished by standardizing the feature values, followed by calculating a weighted average of the standardized feature values, and finally by applying a linear transformation to achieve a desired scale (usually by matching some human scoring standards). Standardization can be based on previous parameters for the means and standard-deviations (*SD*) of features. The weights can be based on multiple regression of the features on a human score criterion, but they can also be based on human expert judgments on the importance of features (e.g., see Ben-Simon & Bennett, 2006). The simplicity of this process and the small and standard feature set allows this process to be performed on the fly (Attali, 2006).

Table 1

Features Used in the Present Study

Feature	Description
Grammar	Based on rates of errors such as fragments, run-on sentences, garbled sentences, subject-verb agreement errors, ill-formed verbs, pronoun errors, missing possessives, and wrong or missing words
Usage	Based on rates of errors such as wrong or missing articles, confused words, wrong form of words, faulty comparisons, and preposition errors
Mechanics	Based on rates of spelling, capitalization, and punctuation errors
Style	Based on rates of cases such as overly repetitious words, inappropriate use of words and phrases, sentences beginning with coordinated conjunctions, very long and short sentences, and passive voice sentences
Essay length[a]	Based on number of words
Vocabulary	Based on frequencies of essay words in a large corpus of text
Word length	Average word length

[a]A measure of essay length was used in this study instead of the organization and development features of *e-rater* V.2.

Attali and Burstein (2006) evaluated the true-score correlations between human essay scores and *e-rater* V.2 scores in the context of K–12 writing. This kind of evidence is important because it directly estimates the overlap in systematic variation of human and automated essay scores. In the context of Messick's (1989) framework for validity, these correlations are related to two different aspects of the construct validity of AES. To the extent that AES are viewed as a qualitatively different method of essay scoring than human holistic scoring, these correlations support the external aspect of construct validity, in the tradition of the multitrait, multimethod approach. However, as AES will be accepted and used as just another rater, these correlations will support the generalizability aspect of construct validity, or in other words, the reliability of scores across different essay raters.

By examining multiple essay scores by the same students, Attali and Burstein (2006) were able to estimate the alternate-form reliability of automated and human scoring and to correct the raw correlations between scores across the different essays. They found that *e-rater* scores had a higher reliability than a single human score (.59 vs. .51) and that the true-score correlation between human and *e-rater* scores was very close to unity (.97). Attali (2007) found similar results for Test of English as a Foreign Language™ (TOEFL®) essay writing by English-as-second-language (ESL) students. The test-retest reliability of *e-rater* scores was .71, whereas the reliability of a single human rater was .54 and that of two human raters was .63. The true-score correlation between *e-rater* and human scores was again .97.

The Present Study

With traditional human essay scoring, or with AES that is calibrated to topic-specific scoring standards, maintaining the comparability of scores across different topics of an assessment is a continual challenge. Comparing essay scores across assessments is practically impossible because there is no satisfactory way to translate the subjective interpretation of scoring guides without actually scoring essays from one assessment under the scoring guide of another assessment.

The advances of *e-rater* V.2 may be used to do just that: enhance the comparability of essay scores across assessments. This is because it is possible to develop a single scoring model across writing assessments, thus using the same scoring standards to evaluate essays written for different assessments. The greatest potential in comparing essays across assessments is in allowing comparability of writing performance along the developmental continuum. A

developmental writing scale could serve both as a tool of educational policy to monitor and evaluate educational progress and as an instructional tool for teachers and students.

In this study, a developmental scale from Grade 4 to Grade 12 for timed essay writing performance was devised on the basis of the *e-rater* V.2 features. Data for the development of the scale was gathered from a national sample of 170 schools representing more than 500 classes from 4th, 6th, 8th, 10th, and 12th grades and more than 12,000 students. Student and school background information was also collected. The students wrote (in 30-minute sessions) up to four essays on topics selected from a pool of 20 topics designed for the corresponding grade levels. The data collection design allocated each topic to classes in up to three grade levels (e.g., 4th, 6th, and 8th grades) in order to allow greater comparability across grade levels. The scale was based on two modes of writing, descriptive and persuasive, and students wrote up to two essays in each mode of writing.

For measurement reasons, the organization and development features in *e-rater* were replaced with a measure of fluency based on essay length (the natural logarithm of the number of words in the essay). The squared multiple correlation of the organization and development features in predicting essay length was very high (around .95), and the correlation between organization and development was found to be negative (the only negative correlation between *e-rater* features). This finding suggests that the two features represent a decomposition of essay length and do not contribute significant unique information to the measurement of essay writing. Essay length itself is the most important objective predictor of human holistic essay scores and exhibits very high test-retest reliability (Attali & Burstein, 2006). Thus, in this study, seven features were used to produce essay scores.

The weighting of features for the computation of scale scores in this study was based on an approach that relies solely on the essay data. The weights were derived from a factor analysis with a single factor. In factor analysis, only the shared variance between variables (features in our case) is analyzed, and an attempt is made to estimate and eliminate variance due to error or variance that is unique to each variable. The derived factor weights are similar to standardized multiple-regression weights in that they estimate the relative contribution of each variable to the common variance among all variables. Thus, the scale scores in this study reflect the relative importance of each feature to the underlying common factor among them.

This approach represents a major departure from the usual approach of basing AES on the prediction of human scores. However, it fits the goal of developing an objective performance measure for writing. This goal also determines to some extent the decision to compute overall writing scores using a single-factor solution and weights. Although this solution is certainly supported by the data (the first factor accounted for 82% of the variance in observed feature values), it is also more convenient to compute overall essay scores that reflect human holistic scores. However, the use of factor analysis opens additional possibilities for computing subscores. These more complex characterizations of essay performance are explored, through exploratory and confirmatory factor analysis, in the later parts of the paper.

The major research question of this paper is the tenability of the assumption of a single scale from Grade 4 to Grade 12. This assumption implies that it is possible to compare the writing performance of students from very different grade levels by using the same scoring standards. It implies that a similar construct underlies the writing performance of children from late elementary school up to the end of high school. It means, for example, that we can assign the same topic to students of different grades and create meaningful score comparisons for students from these different grade levels.

On the face of it, the cognitive literature on the development of writing in children is not incompatible with a single developmental scale. The *knowledge telling* model of writing (Bereiter & Scardamalia, 1987) applies to children from the early phases of writing to adults. The processes that affect performance are shared by younger and older children, and prerequisite transcription processes (spelling and handwriting) typically reach sufficient fluency by fourth grade.

Preliminary evidence for a single developmental scale also comes from Attali and Burstein (2006). Their results show that (a) all of *e-rater*'s features predict human essay scores for essays written by children from 6th grade to 12th grade, (b) in all of these grade levels the true-score correlations between automated and human scores are very high, and (c) the relative weights of the different features in the scoring models do not vary significantly across grade levels. This evidence indicates that the same small set of objective measures of writing performance accounts for the variability in human essay scores across a large range of grade levels.

A major problem in the construction of developmental scales (in mathematics, for example) is the fact that knowledge differs markedly across grade levels. Even a very high achiever in fourth grade would not be able to solve math problems designed for seventh grade

because the seventh-grade curriculum includes material that the fourth grader was not exposed to. This complicates, and limits, the interpretation of the developmental scale. The advantage of a writing developmental scale based on topic-assigned essay writing is that the content of items, the topic texts, is similar across grade levels. In this study, the actual range of cross-grade topic assignments was 4 years, or five grade levels (from 4th to 8th grade, 6th to 10th grade, or 8th to 12th grade).

Several kinds of analyses were performed in order to validate the developmental scale. Multilevel analyses at the essay, student, and topic levels were performed to estimate how overall variability of scores was distributed across these levels. In the context of a standardized assessment, different topics are supposed to be interchangeable. Thus, significant differences in mean scores across topics compromise the validity of the developmental scale.

The multilevel analyses were also intended to estimate the effects of important essay-, student-, and topic-related variables on scores. Variables that affect scores in relation to the validation of the developmental scale include essay order, mode of writing, and topic grade level. Since little or no learning could have occured in the short time period the essays were written, we expected small differences in performance between essays. Topic mode might have been an important predictor of cross-topic variance. Since every topic was originally designed for one specific grade level, it was important to confirm that presenting topics to different grade levels (e.g., an eighth-grade topic to sixth-grade students) did not affect their scores. Finally, these analyses provided an opportunity to evaluate the differences in scores across gender, ethnicity, English language background, and focus on writing in school.

A second validation effort was conducted as an experiment where human raters scored the essays written to two topics (one descriptive and one persuasive) by students in 6th, 8th, and 10th grades. The raters scored half of the essays under the assumption that these were written by sixth graders, using a sixth-grade scoring guide. The other half of the essays was scored under the assumption that these essays were written by 10th graders, using a 10th-grade scoring guide. In fact, both samples of essays included essays from all three grade levels. The purpose of the experiment was to see whether the real grade level of the students had a different effect on automated scale scores than on human scores. In particular, one possibility was that the maturity of writing would influence human scores in ways that the automated scores would not be sensitive to. This possibility would imply that under the 6th-grade scoring model, the human

scores of the 10th-grade students would be higher than their automated scores, and under the 10th-grade scoring model, the human scores of the 6th-grade students would be lower than their automated scores. The existence of this interaction would undermine the validity of the developmental writing scale.

A third validation effort was conducted by collecting longitudinal data on a sample of participating students across the 2 years of data collection. The main study and the developmental scale scores were based on cross-sectional data. The longitudinal data allowed a direct test of the model predictions for the advance in performance that the students were expected to make across a 1-year time span.

The fourth validation effort involved a factor analysis that was conducted on the data at each grade level. The purpose was to explore the underlying structure of the features to find out how similar this structure was across grade levels. Since such analyses of objective writing features do not exist in the literature, extensive exploratory factor analyses were conducted first, followed by confirmatory factor analyses of several structures.

These four validation efforts provided a diverse set of evidence about the feasibility of the developmental scale. The paper concludes with a discussion of some possible applications of the developmental writing scale.

Method

School Sampling Plan

In this section, the sampling plan for schools invited to the study is briefly reviewed. A more elaborate description of the plan is presented in the appendix. The overall plan was to recruit a national sample of schools that would contribute at least one class for each grade included in the study (4th, 6th, 8th, 10th, and 12th grades). The sampling of invited schools was based on information extracted from the latest available (October 2004) comprehensive national school surveys prepared by the National Center for Education Statistics (NCES; NCES, 2006a' NCES, 2006b). School background information included in these surveys was used to ensure the representativeness of the invited schools' sample. The variables used were school region, location, percent of students eligible for free or reduced-price lunches, percent of minority students, school size, and school type (public or private).

Topic Selection and Allocation to Classes

The purpose of the topic allocation plan was to assign topics to two or even three adjacent grade levels (e.g., a particular topic would be presented to 6th, 8th, and 10th graders) in order to better link writing performance across grades.

Twenty topics were selected for this study from the *Criterion*SM topic pool [1] (the list of topics in this pool can be accessed in the *Criterion* Web site, criterion.ets.org). In this pool, topics are categorized by grade and mode of writing. For this study, selected topics were also categorized in this way, and two topics (labeled A and B) were selected for each combination of grade (4th, 6th, 8th, 10th, and 12th) and mode (persuasive and descriptive). The topics were labeled according to the above classification, for example, PA4 is the fourth-grade persuasive prompt A. In five cases, it was necessary to assign a study topic for a specific grade (e.g., an eighth-grade topic) from an original *Criterion* topic of an adjacent grade (e.g., a ninth-grade topic).

Classes were randomly assigned to one of several sessions defined by the identity and order of topics students were asked to write about. Each class was assigned two topics per mode. Each one of the two topics in each mode was either at the same grade level as the class grade (e.g., a sixth-grade topic for a sixth-grade class), or at an adjacent grade level (e.g., either a fourth- or eighth-grade topic for a sixth-grade class). Three combinations of class grade with topic grade were used: one topic grade below grade level and one at the same grade level, one topic grade above grade level and one at the same grade level, or one topic below grade level and one above grade level. An example of the first combination would be a sixth-grade class assigned one fourth-grade topic and one sixth-grade topic. Naturally, 4th grade and 12th grade classes could be assigned to only one of these combinations. In addition, in the second year of the study, only the first two combinations were used.

Several additional constraints defined class sessions: a class could be assigned either the two persuasive or the two descriptive topics first with the two within-mode topics always assigned consecutively (resulting in two combinations); a class could have been assigned either the A or B topics (again resulting in two combinations); and the internal order of within-mode topics also could have been varied (again resulting in two combinations).

Overall, these combinations created 24 different topic sessions for 6th, 8th, and 10th grade classes (16 in the second year, when the below-above combination was dropped), and 8

different topic sessions for 4th and 12th grade classes. Table 2 presents all session definitions for sixth graders. Classes were randomly assigned to one of these sessions. For mid-grade classes, a particular topic (from the same or adjacent grade level) was presented to one third of the classes. For example, sixth graders were assigned to topic PA8 (the eighth-grade persuasive A topic) if their session was defined by A topics (half of sixth-grade classes) and either the below-above or at-above combinations (two thirds of classes).

Table 2

Possible Topic Assignments for Sixth-Grade Classes

		Below/At grade level				At/Above grade level				Below/Above grade level			
A or B	Mode order	1	2	3	4	1	2	3	4	1	2	3	4
A	D-P	DA4	DA6	PA4	PA6	DA6	DA8	PA6	PA8	DA4	DA8	PA4	PA8
		DA6	DA4	PA6	PA4	DA8	DA6	PA8	PA6	DA8	DA4	PA8	PA4
	P-D	PA4	PA6	DA4	DA6	PA6	PA8	DA6	DA8	PA4	PA8	DA4	DA8
		PA6	PA4	DA6	DA4	PA8	PA6	DA8	DA6	PA8	PA4	DA8	DA4
B	D-P	DB4	DB6	PB4	PB6	DB6	DB8	PB6	PB8	DB4	DB8	PB4	PB8
		DB6	DB4	PB6	PB4	DB8	DB6	PB8	PB6	DB8	DB4	PB8	PB4
	P-D	PB4	PB6	DB4	DB6	PB6	PB8	DB6	DB8	PB4	PB8	DB4	DB8
		PB6	PB4	DB6	DB4	PB8	PB6	DB8	DB6	PB8	PB4	DB8	DB4

Note. A or B = prompt; D = descriptive; P = persuasive; PA4 = fourth-grade persuasive prompt A.

School Invitation

From November 2004, about 12,000 public schools and 900 private schools were sent invitations to participate in the study. Overall, about 280 schools responded (a little over 2%) with interest in participating. Finally, 134 schools participated in the first-year study, constituting about 1% of invited schools.

Around November 2005, 4,000 additional public schools were invited to participate in the second-year study. Only new public schools were invited in the second year because the participation of private school students in the first year was higher than expected. A total of 58

new schools responded to these invitations (about 1.5%) with interest in participating. Finally, 34 new schools participated in the second-year study, constituting about 0.8% of newly invited schools.

In addition to the new schools that were invited to participate in the second-year study, schools that participated in the first year were invited to participate again and were encouraged to register actual classes that had participated in the first year and were now in odd grade levels (5th, 7th, 9th, or 11th grades). This invitation created a follow-up sample of students that would allow some validation of the cross-sectional results through a longitudinal study perspective. A total of 25 such schools participated in the second-year study, and 15 of them contributed odd grade-level classes to the study.

Finally, a few schools that expressed interest in the first-year study but did not eventually participate were also invited again, and two such schools participated in the second year.

Procedures

Schools that were interested in participating were asked to complete the school contact information and information on classes that would participate, including grade level, teacher name, and approximate number of students. Schools were limited to a maximum of four classes per participating grade level. In the first year of the study, the invitation letter also included a class registration sheet. In the second year, a Web site registration page replaced the sheet.

Classes were randomly assigned to a particular topic session. For each school a unique school identifier was created, and for each registered class a school-unique class identifier was also created (based on the teacher's last name). These identifiers were sent (by e-mail or fax) to the school's contact person, together with detailed information on the procedures for student registration and use of the study Web site. All classes were assigned a fixed set of 50 unique student registration codes (combination of three letters). Teachers were to assign a code to each student (and keep a record of these assignments in case students forgot their code).

Teachers were encouraged to login to the study Web site in advance of the first class session to get acquainted with the procedures for registering students and writing essays. In the first screen of the study site, the student was asked to enter the school and class identifiers. A second login page followed in which the student was asked to enter his or her personal three-letter code. In case a student in this class did not yet use this code, the student was asked to complete a background questionnaire (more on this below). If the code was already in use, a

third login page appeared in which the student was asked to enter his or her first name, as it was entered in the background questionnaire.

Next came several tutorial pages that explained how to write, edit, and submit the essay. This tutorial was followed by the essay writing page, with the text of the topic and general instructions present at all times. The students were given 30 minutes to complete their essays. It was possible to save an unfinished essay without submitting it and return to it at a later time. The essay text was also regularly saved in case of a technical failure. The order of the topics that were presented to students followed the topic session order that was assigned to their class.

When the essays were submitted, they were immediately sent for automatic scoring by *e-rater*. After a few seconds, a score and feedback report was created for the student. The *e-rater* essay score was based on the *Criterion* scoring model for the grade level of the student. The feedback report was similar to the *Criterion* report that students receive after submitting essays (a demo tour of the *Criterion* system and the essay feedback provided is available in http://www.ets.org/Media/Products/Criterion/tour_06). The reports were saved in the system and both students and teachers could access and print a version of the report at any time.

Technical support for accessing and using the study Web site was available for the study participants from the ETS technical support group and from the study administration staff via phone and e-mail.

Background Questionnaire

The background questionnaire that students completed before writing their first essay included their first name, last name (optional), date of birth, gender, and ethnicity/race (optional). The students were also asked what was the first language they spoke (English, English and another language, or non-English); what grade they started attending a school in which instruction was in English; and how much emphasis was placed on writing in English or language arts classes (hardly any time, a small amount of time, a fair amount of time, or most of the time).

Results on the Participation of Classes and Students

Overall, 527 classes and 11,955 students participated in the regular study (even grade-level classes) during both years, with an average class size of around 23 (M = 22.7, Mdn = 22, SD = 9.5). Although around 90% of the classes had between 10 and 40 students, six classes had only one student and four classes had 50 students, the maximum allowed in this study.

The essays were written between January and June (in both years), and most of the essays (60%) were written in April and May. The median amount of time for students to complete all their essays was 19 days, and the 90th percentile was 63 days.

Table 3 summarizes, for each grade level, the number of students, classes, schools, and submissions per student. The table shows that the sample included fewer fourth graders than higher grade students and that the average number of submissions per student was around three, out of the maximum four submissions planned in the study. There was also a slight decrease in submissions in higher grade levels.

Table 3

Number of Students, Classes, Schools, and Essays

Grade	Students	Classes	Schools	Mean submissions
4	1,156	57	27	3.1
6	2,214	95	46	3.2
8	3,374	133	66	3.0
10	2,899	128	67	2.9
12	2,312	114	55	2.8
All	11,955	527	170[a]	3.0

[a] Schools overlap across grade levels. Overall, 170 different schools participated.

The main goal of this study was to estimate writing performance at different grade levels. Therefore, it was necessary to estimate the success in representing the population of students at each grade level and, when necessary, to correct biases in the sample relative to the population.

Table 4 shows, for several important factors used in the sampling plan, the overall relative distributions of students in the sample and in the entire population of schools from which schools were invited. More than 17 million students were enrolled in relevant grades (even grade levels) in this population of schools. This number means that the main study sample size is less than 0.07% of the relevant population of students. Notable discrepancies between sample and population distributions are (a) lower than expected number of public school students, (b) lower than expected number of city students, (c) lower than expected students from schools with high percentage of minority students, (d) lower than expected students from public schools with high

percentage of students eligible for free or reduced-price lunches, and (e) lower than expected students from western states.

Table 4

Percentage of Students in Sample and Population by Different Factors

Factor	Sample	Population
School type		
Private	14	9
Public	86	91
Locality		
City	20	30
Urban fringe	49	44
Rural	32	26
Minority		
Low third[a]	32	24
Mid	37	36
High	32	40
Lunch		
Low third[a]	42	38
Mid	34	31
High	24	31
Region		
Central	21	24
North east	30	20
South east	35	23
West	14	33

Note. Minority = percentage of minority students; Lunch = in public schools, percentage of students eligible to receive free or reduced-priced lunches.

[a] Thrity-three and 67 percentile ranks were computed separately for different school types and public school level and did not take into account school size.

In order to correct discrepancies between sample composition and the population of students, a cross-table of expected and observed relative frequencies was created with respect to some of the factors used in sampling. This computation was performed separately in each grade level because estimation of writing performance will be accomplished separately in each grade. Two of the above-mentioned factors were not used in creating the cross-table: region, because it is less important for writing performance; and eligibility for free or reduced-price lunches, because it is restricted to public schools. In each grade, the expected and observed relative frequencies for different school types, localities, and, for public schools only, minority enrollment thirds, were computed. Minority enrollment was not used for private schools because the size of the subgroups formed would be too small. Overall, this categorization created 60 subgroups: 45 for public schools (5 grade levels x 3 localities x 3 minority levels) and 15 for private schools (5 grade levels x 3 localities).

Table 5 shows the observed and expected relative frequencies of some of these 60 subgroups. The subgroups presented are the ones that showed the largest discrepancies in terms of the discrepancy effect size, Cramer's *V*. *V* is derived from the regular χ^2 statistic. However, whereas χ^2 is based on frequencies, *V* is based on relative frequencies. More specifically, for *k* x 2 tables (here *k* equals 12 for each grade level), $V = [\chi^2 / N]^{1/2}$. For example, the first two rows in the table contributed 0.97 and 0.59 to the fourth-grade overall effect size—the two largest discrepancies in the table. Both subgroups were overrepresented in the fourth-grade sample, with almost one quarter of the fourth-grade sample coming from the first group whereas only about 4% were expected. The table also shows that 6 of the 10 more significant discrepancies were associated with fourth-grade subgroups.

Table 6 shows the overall effect sizes (*V*) for each grade level. One can see that the fourth-grade *V* is very large, the sixth-grade *V* can be considered medium (Cohen, 1988), and the other three effect sizes can be considered small (Cohen). However, these discrepancies in themselves do not mean that the sample has produced a biased estimate of writing performance in the population. In order to both assess the bias and to try to correct it, *student weights* were computed for each of the 60 subgroups discussed above by dividing the expected by the observed percentage of students in each group. The purpose of these student weights is to correct within-grade discrepancies by over- or underweighting the results for each student in the sample

according to their subgroup affiliation. In the next section, the effect of these sample corrections on the essay writing scores will be evaluated.

Table 5

Subgroups Contributing More Than 0.04 to Cramer's V Computation

Grade	School type	Locality	Minority	*N*	% expected	% observed	χ^2 / *N*
4	Private	Urban	Low	286	4.3	24.7	0.97
4	Public	City	Low	123	1.4	10.6	0.59
4	Public	City	High	67	17.8	5.8	0.08
4	Public	Urban	Mid	23	16.5	2.0	0.13
4	Public	Urban	High	10	12.4	0.9	0.11
4	Public	Rural	Low	294	11.9	25.4	0.15
6	Private	Urban	Low	367	4.2	16.6	0.37
8	Public	Rural	Low	655	11.8	19.4	0.05
10	Public	City	High	220	19.2	7.6	0.07
10	Public	Urban	Mid	814	17.6	28.1	0.06

Note. Minority = percent of minority students.

Table 6

Cramer's V Effect Sizes for the Discrepancy Between Sample and Population

Grade	*V*
4	1.44
6	0.68
8	0.41
10	0.47
12	0.39

Creating Essay Writing Scores

A crucial advantage of developing grade-level norms of writing performance based on objective and automatically computed measures is the possibility to develop a single scoring standard across grades and writing topics. To realize this advantage, all essays should be scored in the same way, regardless of topic, student grade, or mode of writing. Following the terminology of scoring model development in *e-rater* V.2, this entails the standardization of all feature scores, the combination of the standardized feature scores using a relative feature weighting scheme, and finally scaling the weighted standardized scores in some convenient way.

In order to standardize feature scores, all essay submissions in the main study were used. Then a feature weighting scheme had to be adopted. This issue was crucial in the interpretation of scores for this study. Different considerations can guide the development of a feature weighting scheme. One possibility is to base weighting on content expert views in this matter. Ben-Simon and Bennett (2006) asked panels of writing teachers and other content experts to weight the different features of *e-rater* V.2, and Attali (2006) developed an interactive system to allow teachers to customize an *e-rater* scoring model on the fly. It was unclear how experts would respond to the request to develop a single weighting scheme across different grade levels.

Another possibility is to base the feature weighting scheme on empirical results in the development of *e-rater* scoring models based on predicting human scores of essays from different grades. Attali and Burstein (2006) provide evidence that optimal weights are similar across grade levels (Table 4), with typically higher weights for the organization and development features. Table 7 shows a *pseudo-optimal* weighting scheme obtained from a single *e-rater* model for predicting the human scores for a dataset of about 7,600 essays (the same dataset discussed in Attali and Burstein, 2006) written by 6th to 12th graders on 36 topics (about 210 essays per topic).

Another possibility for weighting features that was explored by Attali (2007) is the equal-weights scheme. In this study, the construct validity of the optimal-weights scheme was compared with that of the equal-weights scheme, in the context of TOEFL essay writing. Results showed that the equal-weights scheme was as reliable and showed several advantages over the optimal-weights scheme, such as lower correlations with essay length (due probably to the lower weights for organization and development) and better alignment with the factor-analysis structure of the data.

Table 7

Alternative Feature Weighting Schemes

Feature	Equal weights	Pseudo-optimal weights	Factor weights
Organization	12.5%	34.1%	
Development	12.5%	21.4%	
Essay length			30.8%
Usage	12.5%	11.5%	8.5%
Grammar	12.5%	8.9%	18.8%
Vocabulary	12.5%	7.4%	9.5%
Mechanics	12.5%	6.6%	8.2%
Word length	12.5%	5.4%	9.8%
Style	12.5%	4.7%	14.4%

The three alternatives reviewed above rely on different sources of information for developing a weighting scheme: expert human judgment, prediction of human scores, and a default scheme that assumes no information on features. The weighting scheme adopted in this study is based on a different source of information: the internal relationships of the feature values in the data. Using factor analysis with a single factor, the importance of the different features was estimated from the matrix of observed feature value correlations. The initial eigenvalues of the correlation matrix between features showed that the first value accounted for 82% of the variance. In other words, the first factor accounted for most of the variance in observed feature values, across grade levels. A more detailed account of the factor analysis is presented in a later section. However, Table 7 shows the standardized scoring coefficients of the first factor across all essays. These coefficients were used as standardized weights in the computation of essay scores.

The main differences between the factor analysis weights and the pseudo-optimal weights for predicting human scores are a lower weight for the essay length feature compared to the combined weight of organization and development (and consequently higher weights for the rest of the features) and higher weights for the grammar and style features.

By using the feature distributions and the factor-weighting scheme, *weighted standardized scores* were computed as the sum of the product of the standardized feature scores and the feature relative weight. Since the squared multiple correlation of the features with the

single factor was .847, this was also the variance of the weighted standardized scores. The final *scale scores* were obtained by scaling the weighted standardized scores to have an overall observation-weighted mean of 0 and standard deviation of 1.

In addition to feature weighting, a weighting of essay scores was also applied based on the combination of student weights (discussed in the previous section) and on grade weighting. Student weighting corrected within-grade discrepancies in samplings of students. Grade weighting took into account the different number of students across grades in order to equalize the contribution of each grade level. Thus, independently of the student weights, grade weights increased the weights of 4th-grade observations relative to 10th-grade observations. The two weights (student and grade) were combined (by multiplication) to arrive at a combined weight for each observation.

The Writing Scale and Grade Norms

The developmental writing scale and grade norms that follow are presented below. These are presented only for participating, even grade levels. Interpolation is needed to express odd (5th, 7th, 9th, and 11th) grade norms. This process is described in the longitudinal study below.

Table 8 presents descriptive statistics for the scale scores. The table shows that the effect of student weighting on grade-level performance was negligible, for both means and SDs. This is another indication that the representativeness of the student sample in terms of school characteristics was reasonable. The table also shows an increasing degree of negative skewness and positive kurtosis for higher grades. The reason for this is a small number of low scores in the higher grade levels. This effect can be seen more clearly in Figure 1, a histogram of the scaled scores by grade, together with the fitted normal distribution. The increase in average performance is 0.67 between Grades 4 and 6, and drops to 0.24 between Grades 10 and 12. Table 8 also shows a slightly lower variability in Grade 4 and Grade 12 compared to other grades.

Figure 2 compares the idealized normal distributions of scaled scores across grade levels, based on weighted means and SDs. For example, a scaled score of zero corresponds to the 17th percentile in Grade 12, the 25th percentile in Grade 10, the 44th percentile in Grade 8, around the 69th percentile in Grade 6, and around the 92nd percentile in Grade 4.

Table 8

Descriptive Statistics for Scaled Scores

Grade	*N*	Mean	*SD*	Skewness	Kurtosis	Weighted mean	Weighted *SD*
4	3,346	-1.04	0.73	-0.25	-0.13	-1.06	0.73
6	6,837	-0.40	0.79	-0.32	0.07	-0.39	0.80
8	9,967	0.13	0.80	-0.65	0.45	0.15	0.79
10	8,169	0.54	0.79	-0.86	1.11	0.51	0.82
12	6,312	0.75	0.78	-0.93	1.47	0.75	0.76
All	34,631	0.12	0.95	-0.47	-0.14	0.00	1.00

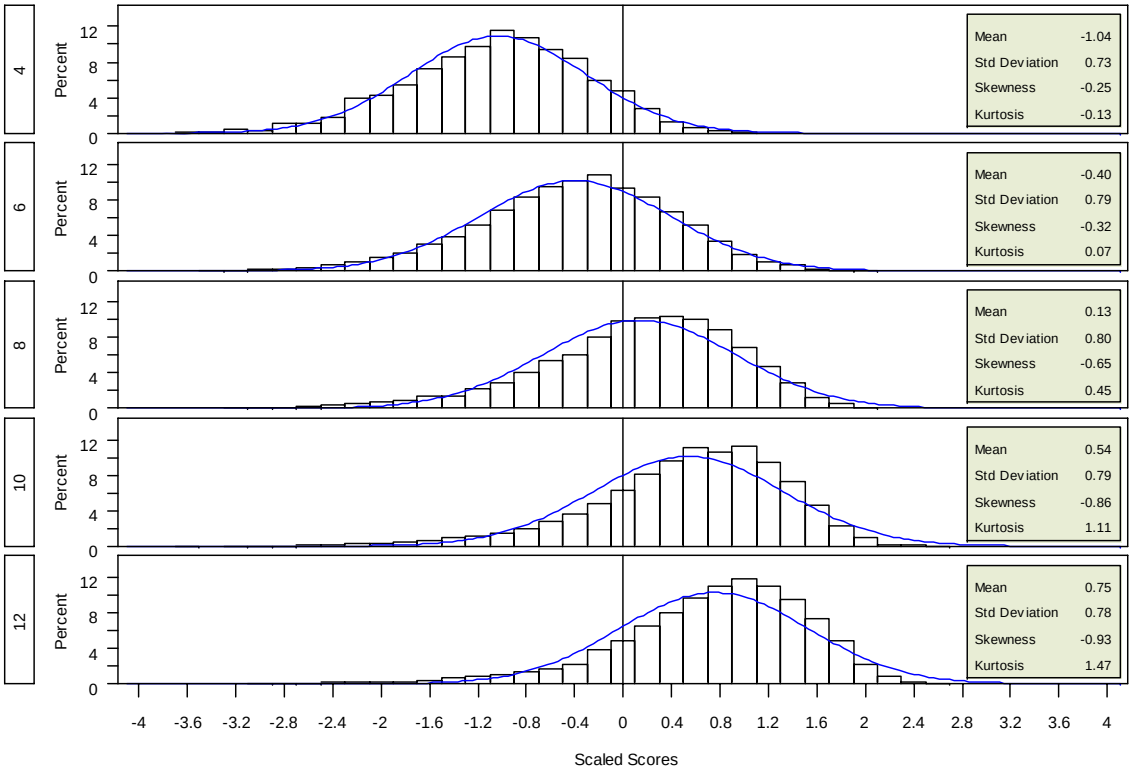

Figure 1. **Histogram of grade norms.**

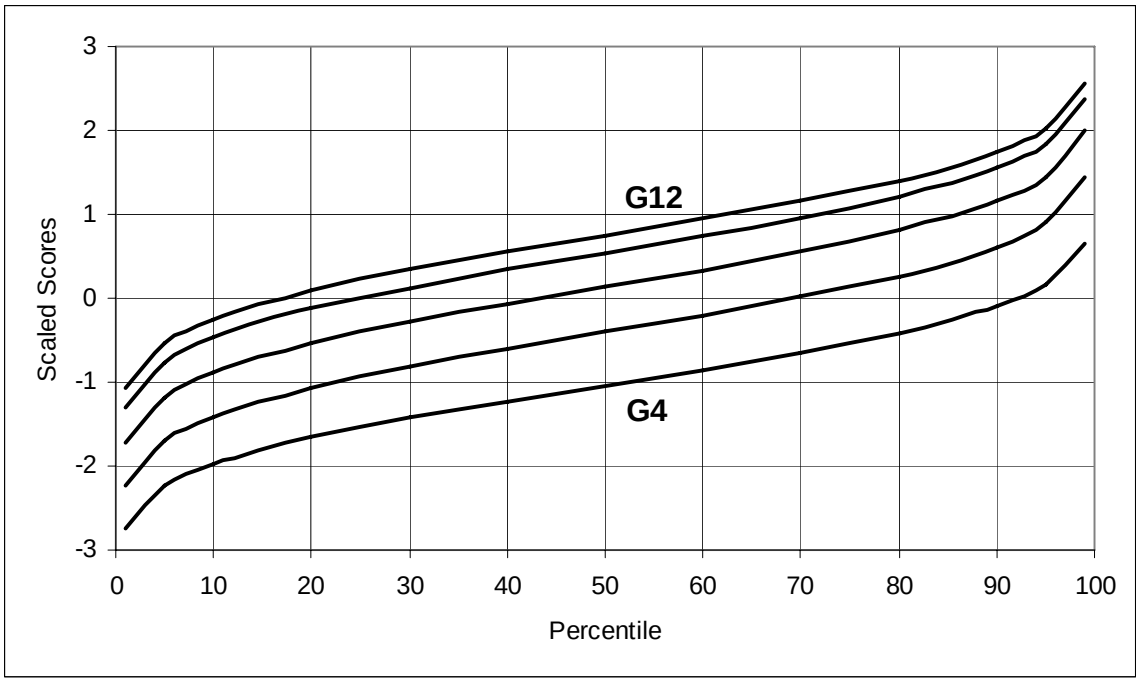

Figure 2. **Normal percentile ranks for grades norms.**

Cross-Classified Random Modeling for Student and Topic Effects

The purpose of these analyses was to model the effects of student and topic characteristics on essay scores in order to partition the variance in essay scores between students and topics and to explain variation in essay scores through different predictor variables. In modeling the student and topic effects, we had to take into account the fact that these two factors did not form a balanced two-way design of the classical analysis of variance. Sample sizes for each topic varied and students wrote essays on only a small subset of the topics. Table 9 shows a summary of the percent of essays written on each topic in each grade level. In this data structure, the lower-level essay scores are cross-classified by two higher level factors—students and topics—and these two factors are treated as random effects (Raudenbush & Bryk, 2002). Modeling of these effects was performed with HLM 6.02 statistical software (Raudenbush, Bryk, & Congdon, 2004). In these analyses, observations were not weighted on the basis of student and grade weights that were used in the preceding section.

The first goal of the analysis was to partition the variance in essay scores between students and topics. The second goal was to explain variation in essay scores through essay (Level 1) as well as student and topic (Level 2) variables. In the first, unconditional model, no explanatory

variables were included; it was used to estimate the baseline variance components of the student and topic factors. The second model introduced the major student level explanatory variable: grade level. In addition to estimating its effect on scale scores, the conditional (on grade level) student variance component in this (more restricted) model could be compared to the unconditional variance component in the previous model to estimate how much of the student variance was explained by grade level.

Table 9

Percentage of Essays Written on Each Topic in Each Grade Level

Topic	G4	G6	G8	G10	G12	Overall
D04A	1.49	1.16	0.00	0.00	0.00	2.64
D04B	1.21	1.24	0.00	0.00	0.00	2.45
D06A	1.40	2.13	1.97	0.00	0.00	5.49
D06B	1.13	1.98	2.17	0.00	0.00	5.28
D08A	0.00	1.48	2.72	1.71	0.00	5.92
D08B	0.00	1.47	2.72	1.84	0.00	6.03
D10A	0.00	0.00	1.87	2.34	2.24	6.45
D10B	0.00	0.00	2.89	2.30	2.41	7.60
D12A	0.00	0.00	0.00	1.97	2.25	4.22
D12B	0.00	0.00	0.00	1.48	2.35	3.83
P04A	1.19	1.32	0.00	0.00	0.00	2.51
P04B	0.94	1.37	0.00	0.00	0.00	2.32
P06A	1.26	2.17	1.99	0.00	0.00	5.41
P06B	1.05	2.17	2.01	0.00	0.00	5.23
P08A	0.00	1.78	2.59	1.84	0.00	6.21
P08B	0.00	1.49	3.00	1.89	0.00	6.38
P10A	0.00	0.00	1.90	2.37	1.95	6.22
P10B	0.00	0.00	2.95	2.21	2.63	7.79
P12A	0.00	0.00	0.00	1.92	1.97	3.89
P12B	0.00	0.00	0.00	1.71	2.44	4.15
All	9.66	19.74	28.78	23.59	18.23	100.00

This process of adding explanatory variables, estimating their effect on scale scores, and estimating how much of the student or topic variance was explained by these variables was repeated with two other classes of variables. The first class included one essay-level and one topic-level variable. The essay-level predictor that was used in this analysis was essay order (first to fourth), and the topic-related predictor was the mode of the topic (descriptive or persuasive).

In the last model, student-related background variables were added, mainly to estimate their effect on scale scores. Student ethnic background (White, Black, Hispanic, and Asian), gender, English as a second language, number of years in English-speaking school, and emphasis on writing in English and language arts classes were also used as predictors of essay scores. In the last two models, several interactions between grade level and other variables were also considered. Descriptive statistics for the variables involved in the analyses are presented in Table 10.

Table 10

Descriptive Statistics for the Cross-Classified Random Model Analyses

Variable	Mean	*SD*
(a) Outcome (*n* = 34,630)		
Scale scores	0.12	0.95
(b) Essay level (*n* = 34,630)		
Essay order (0-3)	1.21	1.09
(c) Student level (*n* = 11,856)		
Grade (-2 = 4th, …, 0 = 8th, …, 2 = 12th)	0.25	1.23
Female (1 = F, 0 = M)	0.53	0.50
White (1 = W, 0 = non-W)	0.70	0.46
Black (1 = B, 0 = non-B)	0.13	0.34
Hispanic (1 = H, 0 = non-H)	0.07	0.26
Asian (1 = A, 0 = non-A)	0.04	0.19
ESL (1 = ESL, 0 = non-ESL)	0.15	0.36
English schooling (2 = Less than one year, 1 = More than a year, 0 = Always)	0.31	0.66
Time for writing (0 = Hardly any, 1= Small amount, 2 = Fair amount, 3 = Most of the time)	1.81	0.72
(d) Topic level (*n* = 20)		
Persuasive (P = 1, D = 0)	0.50	0.51

Note. ESL = English as second language.

The first, unconditional model is used to decompose the essay score variance. At Level 1, the model is

$$Y_{ijk} = \pi_{0jk} + e_{ijk}, \qquad e_{ijk} \sim N(0, \sigma^2),$$

where Y_{ijk} is the *i*th essay score written by student *j* on topic *k*; π_{0jk} is the expected score for student *j* on topic *k*; and σ^2 is the within-cell variance.

At Level 2, the model is a *main effects model,* with the random effects associated with students and topics. The interaction between students and topics is not modeled because, at most, only one observation exists per cell (no student wrote more than one essay for a topic), therefore it is impossible to disentangle the student-by-topic variance from the within-cell variance. Thus, the model is

$$\pi_{0jk} = \theta_0 + b_{00j} + c_{00k}, \qquad b_{00j} \sim N(0, \tau_{b00}), \qquad c_{00k} \sim N(0, \tau_{c00})$$

where θ_0 is the grand mean of essay scores; b_{00j} is the random main effect of student *j*, that is, the contribution of student *j* averaged over all topics, assumed normally distributed with mean 0 and variance τ_{b00}; and similarly c_{00k} is the random main effect for topic *k*.

The results of this model are presented in Table 11. The table shows that most of the variance, 74%, is accounted for by students. Only a small amount of variance, 4.1% of the total, is explained by topics. This proportion of .04 can also be interpreted as the intraclass correlation for topics, the correlation between scores of two essays written by two randomly chosen students on the same topic. Ideally, this correlation should be 0, but the results show a small amount of clustering of scores within topics.

Although the topic random effect is small, part of it is due to the study design, whereby topics were not presented to the full range of grade levels. This means that the average scores for topics that were presented to lower grade levels will probably be lower than those presented to higher grade levels. Thus, the second model presented is one in which a single student-level predictor—grade level—is introduced, in order to estimate the topic effect within grade level.

In this model, a single fixed-effect parameter for the slope of grade level is added. Thus,

$$\pi_{0jk} = \theta_0 + \gamma_{01}(GRADE)_j + b_{00j} + c_{00k}$$

Table 11

Results for Unconditional Model

Fixed effect	Coefficient	*se*			
Overall essay score mean, θ_0	0.0820	0.0428			
Random effect	Variance component	% variance	*df*	χ^2	*p*-value
Student, var (b_{00j})	0.6335	74.1%	11,855	101,247	0.000
Topic, var (c_{00k})	0.0354	4.1%	19	2,680	0.000
Residual error, σ^2	0.1858	21.7%			

The results are presented in Table 12. The grade coefficient is highly significant and can be interpreted to say that the average gain in scores across two grade levels is around 0.44. As expected, this predictor explains an important amount of variance, both between students and topics. By comparing the variance components across the two models, we can measure the proportion reduction in variance, or variance explained, by grade level. For the student variance, the proportion of variance explained by grade is (.6335 - .4521) / .6335 = 29%. For the topic variance, the proportion of variance explained by grade is (.0354 - .0119) / .0354 = 66%. Incidentally, the portion of score variance that lies between topics after controlling for grade level is even smaller than for the unconditional model, only 1.8%.

Table 12

Results for Grade Model

Fixed effect	Coefficient	*se*	*t*-ratio		
Overall essay score mean, θ_0	0.0008	0.0254			
Grade level	0.4358	0.0062	69.797		
Random effect	Variance component	% variance	*df*	χ^2	*p*-value
Student, var(b_{00j})	0.4521	70.2%	11,854	95,200	0.000
Topic, var(c_{00k})	0.0119	1.8%	18	2,853	0.000
Residual error, σ^2	0.1798	27.9%			

Figure 3 visually shows the small variability in essay scores across topics after controlling for grade level. The figure shows the average grade-centered scale scores across all 20 topics. Grade-centered scores were computed by subtracting the average grade score for the corresponding student from every essay score. The average within-topic grade-centered standard deviation is 0.81 whereas the standard deviation of average grade-centered scores across topics is only 0.11. Incidentally, the figure also shows that persuasive topics are associated with higher scores. This effect will be the focus of the next model.

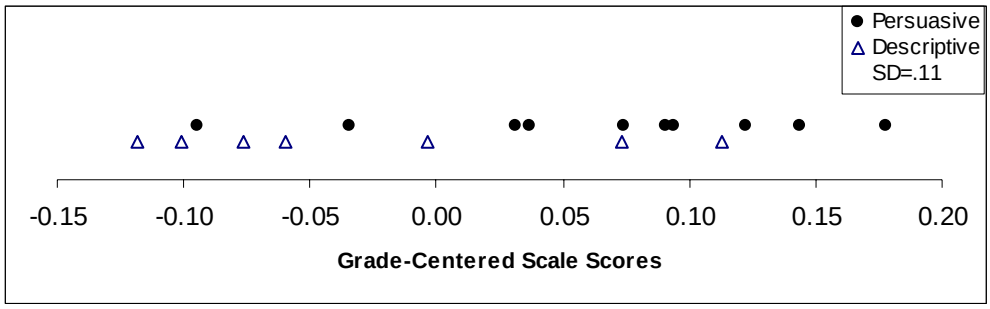

Figure 3. **Grade-centered average scale scores across topics.**

The third model introduces topic-level variables, essay-level variables, and interactions between topic and student variables. The topic-level predictor is mode of writing, to see whether it is possible to explain some of the residual variance between topics. The interaction between student grade and topic mode is also explored to see if the mode effect is different in different grades. Another interaction between student and topic variables is the relative topic grade of the essay—the difference between the original grade level of the topic and the student grade level—is introduced to examine the possible effect of presenting topics that were not intended for the student grade level on performance. Finally, the essay order (first to fourth) as an essay-level predictor is also introduced here.

In the model just described, the relative topic grade variable had a small and nonsignificant effect. Therefore, Table 13 shows the reduced model without this variable. The table shows that the topic effect was further reduced by 52% by incorporating the mode effect ([.0119-.0057] / .0119). All three fixed effects are significant. The mode effect shows that persuasive topics are associated with higher scores (that is, they are easier than descriptive topics), by about 0.14 for

eighth grade first essays (that is, when GRADE = 0 and ESSAY ORDER = 0). Further, the mode effect is stronger for higher grades (a positive interaction with grade) by about 0.04 per two grade levels. This translates into an estimated mode effect for 12th graders' first essay of about 0.22 and an estimated mode effect for fourth graders' first essay of about 0.05.

Table 13

Results for Mode and Essay Order Model

Fixed effect	Coefficient	*se*	*t*-ratio		
Overall essay score mean, θ_0	-0.0782	0.0251			
Grade level	0.4158	0.0070	57.232		
Persuasive mode	0.1366	0.0342	3.999		
Essay order	0.0105	0.0023	4.609		
Grade × persuasive	0.0408	0.0066	6.190		
Random effect	Variance component	% variance	*df*	χ^2	*p*-value
Student, var(b_{00j})	0.4512	70.9%	11,853	95,173	0.000
Topic, var(c_{00k})	0.0057	0.9%	17	1,304	0.000
Residual error, σ^2	0.1795	28.2%			

The essay order effect was also found to be significant, with a small estimated effect (for eighth graders) of 0.01. However, in a Tukey honestly significant difference test (with $\alpha = .01$), only the least-square adjusted means of the first essay were significantly different (lower) than all other subsequent essays. This implies that the essay order effect represents a familiarization effect in the first essay.

The last model introduces all the student level variables and interactions (see Table 14). These variables explain 16% of the student variance (controlled for grade level). Several interesting effects were found. Female students score 0.31 higher on average than male students, but a small negative interaction with grade level (about -0.04 for every two grade levels) was also found. White students score higher than non-White students do (0.11 on average). Black students score lower than non-Black (-0.29 on average), and the effect worsens with grade level (-0.14). Hispanic students do not have significantly lower scores than non-Hispanic students, but there is a

significant negative effect with grade level (-0.07). Finally, Asian students score higher than non-Asian students do (0.39 on average), but again with a negative interaction with grade (-0.07).

Table 14

Results for Student Predictors Model

Fixed effect	Coefficient	*se*	*t*-ratio	*p*-value	
Overall essay score mean, θ_0	-0.3351	0.0407	-8.227	0.000	
Grade level	0.4120	0.0259	15.890	0.000	
Persuasive mode	0.1376	0.0344	4.005	0.000	
Essay order	0.0106	0.0023	4.646	0.000	
Grade × persuasive	0.0408	0.0066	6.199	0.000	
Female	0.3073	0.0127	24.128	0.000	
Grade × female	-0.0395	0.0102	-3.892	0.000	
White	0.1099	0.0277	3.961	0.000	
Grade × white	-0.0186	0.0218	-0.850	0.395	
Black	-0.2934	0.0322	-9.119	0.000	
Grade × black	-0.1350	0.0254	-5.320	0.000	
Hispanic	-0.0246	0.0373	-0.659	0.509	
Grade × hispanic	-0.0736	0.0292	-2.517	0.012	
Asian	0.3904	0.0443	8.820	0.000	
Grade × Asian	-0.0733	0.0336	-2.180	0.029	
ESL	0.0460	0.0255	1.805	0.071	
Grade × ESL	0.0333	0.0178	1.872	0.061	
English schooling	-0.1638	0.0103	-15.864	0.000	
ESL × English schooling	-0.0455	0.0274	-1.663	0.096	
Time for writing	0.0513	0.0091	5.622	0.000	
Grade × time for writing	0.0312	0.0074	4.198	0.000	
Random effect	Variance component	% variance	*df*	χ^2	*p*-value
Student, var(b_{00j})	0.3799	67.2%	11,837	83,044	0.000
Topic, var(c_{00k})	0.0058	1.0%	1	1,276	0.000
Residual error, σ^2	0.1793	31.7%			

Note. ESL = English as a second language.

Surprisingly, ESL students do not earn lower scores on average than non-ESL students (conditional on all other variables in the model). It seems that English schooling is the major linguistic background factor that influences performance (0.16 on average for an increase in one level of English schooling), and there is no apparent interaction with grade. Interestingly, time for writing had a significant beneficial effect on scores (0.05 on average for an increase in one level of time for writing), and the effect grew with grade level (by 0.03 for each two grade levels).

Although the focus of this report is on overall essay scores, it is interesting to examine which features contribute to the mode effect, and thus to the topic effect. Table 15 shows the standardized differences between persuasive and descriptive essays (first essays only) for the seven features used in this study and for the organization and development features of *e-rater* V.2. The table shows that the two word-level features— vocabulary and word length—are responsible for the mode effect. Persuasive topics are associated with less frequent and longer words than descriptive topics, with effect sizes of around 0.5 and 0.8, respectively. Hence, the weights of these two features will have a large impact on the mode effect of essay scores. The table shows how different weighting schemes affect the overall feature-weighted standardized difference (in the last row). When these two features are excluded from the scoring model (the right-most column) and the remaining study weights are proportionally increased, the mode effect disappears.

Score Reliability

The variance components from the cross-classified model results allow us to estimate the reliability of (average) student scores based on essays written to different topics. This is given by

$$\frac{\tau_{b00}}{\tau_{b00} + \tau_{c00} + \sigma^2 / n_j}$$

where nj is the number of essays written by the student. The reliability estimates for the fully unconditional model (Table 11) is .74 for one essay, .83 for two essays, .87 for three essays, and .89 for four essays. Such cross-grade reliabilities are seldom reported, but they do make sense in the context of an assessment that is administered across grade levels. The reliability estimates for the model, conditional on grade level (Table 12), is .70 for one essay, .82 for two essays, .86 for three essays, and .89 for four essays. The reason for the single-essay lower reliability of the

conditional model is the smaller student variance component. However, the smaller topic variance component results in higher reliabilities for student scores based on more essays.

These reliability estimates can be compared to the estimates of human reliability that Breland, Camp, Jones, Moris, and Rock (1987) computed. In a comprehensive empirical study of human scoring reliability, Breland et al. (1987) had students write six topics in three modes of writing, and each essay was rated by three highly experienced readers. Their estimate for a single essay rated by one rater was .42. The estimate for two essays from two different modes, each rated by one rater, was .57 (and .59 for two essays from the same mode). A total of four essays (two for each mode), each rated by a single rater, would reach a reliability of .73, comparable to the grade-conditioned single-essay reliability estimate from this study of .70.

Test-retest reliabilities can also be directly computed from the data. Table 16 presents the estimates for each grade level and overall. The table shows lower reliabilities for Grade 4 and Grade 12. This may be related to the lower standard deviation of scores in these grade levels.

Table 15

Standardized Differences of Feature Values Across Mode of Writing

		Nonstudy weights		Study weights	
Feature	*d*	Pseudo-optimal	Equal	Original	Excluding word-level features
Organization	-0.05	34.1%	12.5%		
Development	0.09	21.4%	12.5%		
Usage	0.11	11.5%	12.5%	8.5%	10.6%
Grammar	-0.01	8.9%	12.5%	18.8%	23.3%
Vocabulary	-0.48	7.4%	12.5%	9.5%	
Mechanics	-0.05	6.6%	12.5%	8.2%	10.1%
Word length	-0.79	5.4%	12.5%	9.8%	
Style	-0.21	4.7%	12.5%	14.4%	17.8%
Essay length	0.09			30.8%	38.2%
Weighted *d*		-0.08	-0.17	-0.12	0.00

Table 16

Median Test-Retest Reliability

Grade	Between mode	Within mode
4	.56	.67
6	.67	.70
8	.67	.73
10	.67	.75
12	.63	.71
Overall	.77	.81

Note. Between mode = median of correlations between Different-Mode Essays 1–3, 1–4, 2–3, and 2–4; Within mode = median of correlations between Same-Mode Essays 1–2 and 3–4.

For the average of two essay scores, the overall reliability estimate within mode (based on the correlation between the average of Essay Scores 1–2 and 3–4) is .85. The overall reliability estimate with confounded modes (based on the correlation between the average of Essay Scores 1–3 and 2–4 and between 1–4 and 2–3) is .90.

Human Scoring Experiment

One of the major premises of this project is that it is possible to automatically score essays written by students attending fourth grade in the same way as essays written by older students attending 12th grade. Correctness of this grade invariance assumption would imply that human raters scoring essays from a diverse range of grade levels would agree with the automated scores of these essays. In other words, the (presumed) higher quality of essays written by older students would be reflected in automated and human scores to the same degree. The rival hypothesis is that automated scores are not sensitive enough to some dimensions of writing that are exhibited differentially in younger and older students. Thus, when automated and human scores will be compared, interactions between scoring mode and grade level will be found. More specifically, if human raters were more sensitive to different aspects of mature writing, we would expect that human scores of older students would increase more dramatically than automated scores.

To test these rival hypotheses, a controlled experiment was conducted, whereby human raters scored essays written by 6th, 8th, and 10th graders according to 6th or 10th grade scoring

standards and without knowing that the essays were written by students from other grade levels. A range of four grades was the largest range of grades for which students wrote essays on the same topics, and the range of 6th grade to 10th grade (even grade levels) was chosen because it was the middle range developmentally.

Students from three grade levels wrote essays for two pairs of descriptive and persuasive topics (these were the original eighth-grade topics). The pair with the most submitted essays (after the first year of the project) was chosen for human scoring. All students from the three grade levels who submitted essays on these two topics were included in this experiment. There were 259 such students from 6th grade, 289 8th graders, and 357 10th graders. For each of these 905 students, two essays—one descriptive and one persuasive—were available for human scoring.

In order to maximize the potential for a scoring mode effect, the raters were told they would score essays from a group of 6th graders and another group of 10th graders. Students in these two grade level groups wrote essays on the same pair of topics. The 10th-grade group would be scored according to the *Criterion* 10th-grade standards, and the 6th-grade group would be scored according to the *Criterion* 6th-grade standards.

In fact, the two groups included students from all three grade levels. The two groups presented as younger and older students were formed by randomly separating each grade-level group into two approximately equal halves. Table 17 presents the design of the two experimental groups.

Table 17

Number of Students in Each Grade Level and Presentation Type

Grade	Presented as 6th graders	Presented as 10th graders	Overall
6	131	128	259
8	146	143	289
10	178	179	357
All	455	450	905

Human scoring was performed by professional ETS raters on a 1–6 scale. The raters went through standard training for the two grade-level scoring standards (6th and 10th grade) using an independent set of essays written by (real) 6th and 10th graders for the corresponding grade-level standards. Each of the essays in the experimental groups was scored by two raters.

Automated scoring was based on the scale scores, which had a different scale than the scale of the human scores. Since we were not interested in this experiment in any overall difference between human and machine scores, the scale scores were scaled to have the same overall mean and standard deviation of the human scores (across all 905 essays), separately for the descriptive (M = 2.8, SD = 1.0) and the persuasive (M = 2.9, SD = 1.0) topics. The median correlation between each of the two human scores and the scale scores (across apparent grade and mode) was .80, compared to the median correlation between the two human scores of .78. That is, the scale score agreement with a single human rater was at least as high as the agreement between the two human raters.

The main research question of this experiment was whether there would be an interaction between the apparent grade (or grade level of scoring standards) and type of scoring (human or automated). To test this question a repeated-measures analysis of variance was conducted with the mode of writing (descriptive and persuasive) and the type of scoring (human and *e-rater*) as within-subjects independent measures and with grade level and apparent grade level as between-subjects independent measures.

This analysis also allowed testing another research question concerning the possible presentation effect on the human scores. Specifically, is it possible that the presentation of students as younger (6th graders) as opposed to older (10th graders) would result in lower human scores.

Figure 4 presents the means of the *e-rater* and human scores for each of the experimental groups. With respect to the main research question about a possible interaction between type of scoring and (true) grade, the figure shows that there is no such interaction in any of the panels corresponding to the different modes of writing and presentation. This interaction was indeed not significant, $F(2, 884) = 0.81$, $p = .44$. Moreover, none of the three-way interactions with mode of writing or grade presentation was significant (and neither was the four-way interaction).

The effect of presentation on scores can be seen in Figure 4 by comparing the same type of scores across the two panels horizontally. One can see that for the human scores the means in

the right panels (for students presented as 10th graders) are indeed higher than the corresponding means in the left panels (for students presented as 6th graders). The overall effect for the human scores is 0.25 points (across the two modes) and for the scale scores only 0.02 points. The interaction between type of scoring and grade presentation was indeed highly significant, $F(1, 884) = 59.4$, $p < .0001$, partial η^2 2 = .03. None of the three- or four-way interactions with grade and mode of writing were significant.

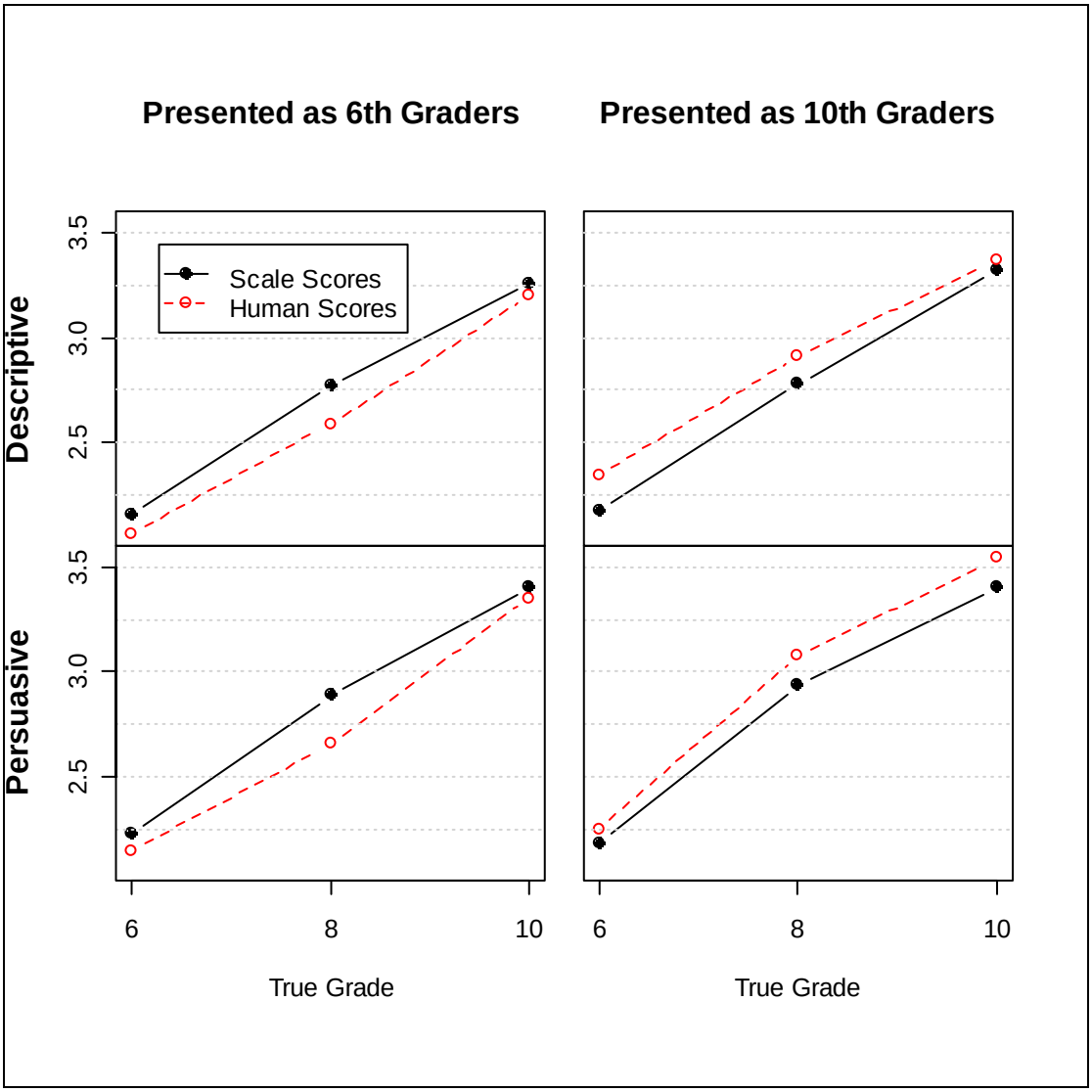

Figure 4. **Profiles of scores for true grade, writing mode, and grade presentation.**

The results of this experiment showed no grade-related bias between human and automated scale scores in scoring 6th to 10th grade essays. This provides further support for the assumption that a single scale can be used to score essays written by students in very different levels of writing proficiency (and maturity). The results also showed what kinds of difficulties would be encountered in trying to develop such a scale based on human scoring since it is difficult for human raters to ignore their knowledge about the grade level of the student writing the essay.

Longitudinal Study

The main study and most of its results are cross-sectional in nature. However, in a further effort to validate the cross-sectional results, a small longitudinal study was undertaken. To accomplish this task, the schools that participated in the first-year project were also invited to participate in the second year. They were encouraged to register both even-grade classes and odd-grade classes, particularly classes that had participated in the first-year study.

Overall, 15 schools contributed 1,380 odd-grade students in the second year of the study. There were 296 students in 5th grade, 608 students in 7th grade, 238 students in 9th grade, and 238 students in 11th grade. Altogether, 401 odd-grade students from 12 schools were identified as repeater students from the first year of the study (29% of the total odd-grade students). Identification was based on last name and date of birth (as reported by the students in both years of the study) together with a close or perfect match on the first name. There were 125 repeater students in 5th grade, 221 repeater students in 7th grade, and 55 students in 11th grade. Unfortunately, no ninth graders were identified among the repeater students.

The purpose of the repeater student analyses was to confirm the expected progress in writing performance over a 1-year period. A useful way to conceptualize the expected results is to characterize the performance of students in relation to their age group. Although individual students can certainly change their relative standings within their age group over time, one can expect that on average the standardized within-grade scores of a large group of students will remain the same from year to year.

One complication in applying this plan for the odd-grade scores is that we have normative performance parameters only for the even grade levels and not for the odd-grade levels. However, one can estimate the mean and standard deviation of scale scores at each odd grade level by averaging the two estimates above and below each odd-grade level. For example,

the mean student-weighted scale scores in fourth and sixth grades are -1.02 and -0.40, respectively (see Table 18). Thus, one can estimate the fifth-grade mean scores by interpolating the fourth- and sixth-grade values. Interpolating can be performed for both the mean and the standard deviation of scores.

Table 18

Actual (Even Grades) and Cubic-Spline Estimates (Odd Grades) for Student-Weighted Score Means and Standard Deviations

Grade	Mean	*SD*
4	-1.064	0.728
5	-0.721	0.772
6	-0.394	0.797
7	-0.099	0.792
8	0.154	0.787
9	0.355	0.807
10	0.513	0.824
11	0.637	0.806
12	0.745	0.765

Interpolation can be done in different ways. One of the simplest is linear interpolation, where midpoints are estimated by connecting two existing adjacent points by a straight line. In polynomial interpolation, the linear interpolant is replaced by a polynomial of a higher degree, with a more accurate estimation result. For n data points, there is exactly one polynomial of degree $n-1$ going through all the points. A more efficient technique is spline interpolation, which uses low-degree polynomials (e.g., cubic) in each of the intervals and chooses the polynomial pieces such that they fit smoothly together. In practice, the differences between the linear and cubic spline interpolations of odd-grade level mean scores were small: -0.004, 0.025, 0.019, and 0.032 for 5th, 7th, 9th, and 11th grades.

Table 18 shows the actual (for even grades) and spline-estimated (for odd grades) student-weighted means and SDs for the scale scores. Using the table figures, all essay scores for repeater students were standardized according to the grade in which the essays were written. The

hypothesis was that the grade-standardized scores of students would not change across the 2 years of their participation.

To test this hypothesis a repeated-measures analysis of variance was conducted with topic, essay order (1-4), year, and the grade students were in during the first year of the study as independent variables, and grade-standardized essay scores as the dependent measure. Two effects were of interest in relation to the main hypothesis: the year effect and the interaction of year with grade in first year. If both are nonsignificant, we cannot reject the hypothesis that the grade-standardized essay scores across the 2 years of participation remained the same. Indeed, both effects were not significant (for the year effect, $F(1, 392) = 0.03$, $p = .87$, and for the year and grade interaction, $F(4, 392) = 1.08$, $p = .37$).

Figure 5 shows the adjusted means for each grade group in the first and second years of participation. Table 19 further shows the actual scale score means in Year 1 and Year 2, and the expected scale score mean in Year 2 (based on the grade-adjusted score mean in Year 1). The figure and table show that the differences in mean scores across the years are small in each of the grade levels (the largest grade-adjusted mean difference, for 10th graders, is -0.11). These results provide further support for the cross-sectional results through a direct longitudinal analysis of student performance.

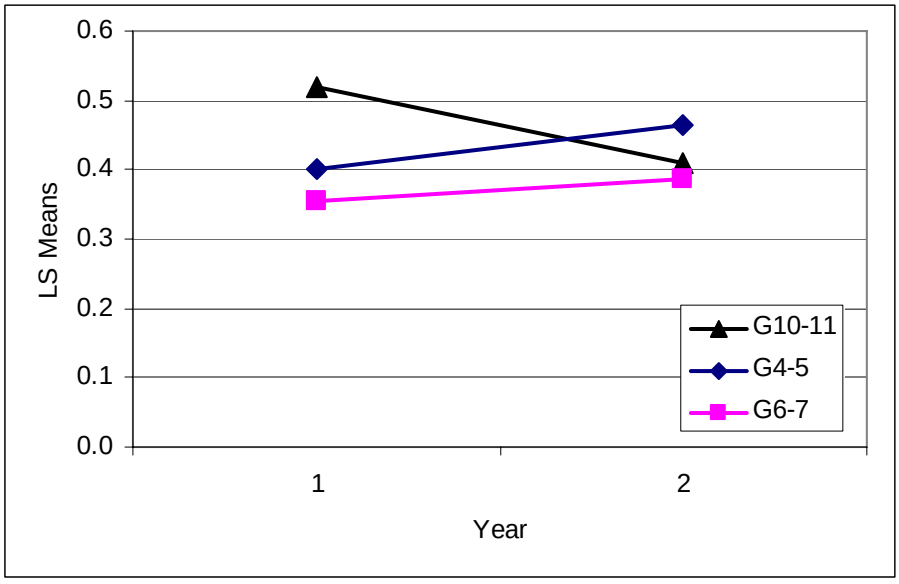

Figure 5. **Least-square means for grade-adjusted new repeater scores.**

Table 19

Longitudinal Study Score Means

	Grade-adjusted scores		Scale scores		
				Actual	Expected
Grade Year 1	Year 1	Year 2	Year 1	Year 2	Year 2
4	0.40	0.46	-0.77	-0.36	-0.41
6	0.35	0.39	-0.11	0.21	0.18
10	0.52	0.41	0.94	0.97	1.06

Factor Analysis

The purpose of this analysis was to offer another kind of evidence on the appropriateness of a single developmental scale by analyzing the internal structure of the different features. To do that, factor analyses were conducted in each grade level to discover which features in the set form coherent subsets that are relatively independent of one another. The similarity of the structure of these subsets (or factors) across grades can support the claim for a single scale.

Table 20 presents the overall correlation matrix for the seven features used in this study. All correlations are positive and range from the teens to the 70s.

Table 20

Overall Feature Correlation Matrix

Feature	S	G	U	M	V	WL
Essay length	.66	.69	.52	.36	.23	.26
Style		.48	.32	.23	.37	.41
Grammar			.51	.43	.20	.23
Usage				.37	.13	.15
Mechanics					.23	.20
Vocabulary						.73
Word length						

Note. $N = 34{,}631$.

Table 21 presents the results of the one-factor analysis across grade levels. The first two rows show the proportion of the first initial eigenvalue and the squared multiple correlation of feature values in predicting factor scores, as an indication of the appropriateness of a one-factor solution. Although these values are high in all grade levels, the table shows a decrease in both measures as grade level increases.

Table 21

One-Factor Solution Across Grades

	G4	G6	G8	G10	G12	All
Proportion of 1st eigenvalue	.87	.86	.76	.74	.69	.82
SMC with factor	.91	.87	.82	.79	.79	.85
Feature communalities						
Essay length	.85	.78	.66	.56	.57	.68
Style	.15	.33	.35	.42	.41	.50
Grammar	.63	.58	.58	.48	.48	.55
Usage	.54	.40	.30	.27	.26	.31
Mechanics	.14	.14	.17	.16	.13	.23
Vocabulary	.01	.01	.05	.14	.13	.20
Word length	.00	.02	.04	.11	.10	.23

Moreover, at the feature level, significant differences are found in communalities (feature variance accounted for by the factors) across grade levels. The communalities of essay length, grammar, and usage decreases; the communality of style increases; and the communalities of vocabulary and word length are low across all grade levels. These results suggest that a higher number of factors could increase communalities of some of the features. It was decided to compare a two- and three-factor solution for interpretability.

Table 22 presents a summary of the oblique-rotated (Promax), three-factor pattern for the seven features across all grades. The table shows the factor number with the highest loading. The features are grouped by their highest loadings on the different factors. The loadings were stable across grade levels, and for most of the features, the same factor was dominant for every grade level. The exception was the grammar and usage features, which were loaded on Factor 1 in lower

grades (4 and 6) and Factor 2 in higher grades (8, 10, and 12). This exception is reflected in a high overall loading of grammar on Factor 1 (.48) in addition to a high loading on Factor 2 (.63).

Table 22

Factor Pattern After Promax Rotation for the Three-Factor Solution

	Factor with loading > .4					All		
Feature	G4	G6	G8	G10	G12	Factor 1	Factor 2	Factor 3
Essay length	1	1	1	1	1	.80	.47	.09
Style	1	1	1	1	1	.68	.20	.32
Grammar	1	1	2	2	2	.48	.63	.09
Usage	1	1	2	2	2	.30	.58	.03
Mechanics	2	2	2	2	2	.09	.57	.16
Vocabulary	3	3	3	3	3	.11	.13	.84
Word length	3	3	3	3	3	.18	.10	.82

An inspection of the two-factor solution (in Table 23) shows that it is similar to the three-factor solution in that the first two factors were merged into Factor 1 in the two-factor solution. However, in the two-factor solution, the loadings of the mechanics feature are low (.37-.38) across all grade levels.

Table 23

Factor Pattern After Promax Rotation for the Two-Factor Solution

	Factor with loading > .4					All	
Feature	G4	G6	G8	G10	G12	Factor 1	Factor 2
Essay length	1	1	1	1	1	.87	.15
Style	1	1	1	1	1	.58	.36
Grammar	1	1	1	1	1	.80	.11
Usage	1	1	1	1	1	.62	.05
Mechanics	-	-	-	-	-	.45	.15
Vocabulary	2	2	2	2	2	.14	.83
Word length	2	2	2	2	2	.16	.85

Table 24 presents the correlations between the Promax-rotated factors for the three-factor and two-factor solutions. The table shows moderate correlations between Factors 1 and 2, low correlations between Factors 1 and 3 and between Factors 2 and 3, as well as between the two factors in the two-factor solution. The correlations are slightly increasing with grade level but quite similar across grade levels.

Table 24

Correlations Between Factors

	3 factors			2 factors
Grade	1–2	1–3	2–3	1–2
4	.38	-.12	-.13	-.11
6	.42	.10	-.08	.11
8	.57	.11	.04	.14
10	.58	.19	.15	.22
12	.59	.18	.07	.19
All	.64	.33	.22	.35

Table 25 shows the final communalities of the features for each solution obtained over all grade levels. The largest gains in increasing the number of factors from one to two are for the vocabulary and word length features. Some gain in increasing the number of factors from two to three can be seen for the essay length and style, as well as for mechanics.

Although the three-factor solution accounts for more observed variance, the previous results suggest that the gains are not uniform across grade levels. The three-factor solution seems more appropriate for the higher grade levels (8, 10, and 12). In lower grades, the two-factor solution seems more appropriate.

A linguistic interpretation of the three-factor solution would regard Factor 1 (see Table 22) as a fluency factor, Factor 2 as a sentence-level conventions factor, and Factor 3 as a word choice factor. This interpretation represents an attractive meaningful hierarchical interpretation from a linguistic point of view. Under this interpretation, it seems that the fluency and conventions factors are not fully distinguished in lower grades, and that the fluency factor is more dominant in these lower grades.

Table 25

Final Communalities for Factor Solutions for Combined Grade Levels

Feature	1 factor	2 factors	3 factors
Essay length	0.68	0.78	0.87
Style	0.50	0.47	0.61
Grammar	0.55	0.65	0.64
Usage	0.31	0.38	0.43
Mechanics	0.23	0.22	0.37
Vocabulary	0.20	0.71	0.74
Word length	0.23	0.75	0.72
Total	2.71	3.96	4.36

This conclusion is also supported by results on the importance of each of the factors. This importance can be assessed by the amount of variance explained by each factor. Table 26 presents the amount of variance explained by each factor after eliminating the other factors (that is, based on the semipartial correlations of the features with the factors). The table shows that in lower grades the fluency factor (Factor 1 in a three-factor solution) is dominant, whereas in higher grades its influence is diminished in favor of the word level factor (Factor 3).

Table 26

Variance Explained by Each Factor Eliminating Other Factors

	3 factors			2 factors	
Grade	1	2	3	1	2
4	1.83	0.64	0.99	2.30	1.04
6	1.60	0.44	1.04	2.22	1.08
8	0.98	0.68	1.30	2.06	1.31
10	0.94	0.65	1.45	1.92	1.45
12	0.90	0.65	1.58	1.91	1.62
All	0.79	0.61	1.30	2.06	1.34

Following the exploratory factor analysis results, multiple-group confirmatory factor analyses were conducted for the five grade levels. First, the alternative models discussed above were tested to determine the number of factors that was supported in the data: one, two (with vocabulary and word length constituting the second factor), or three factors (with grammar, usage, and mechanics separating from the first factor). Based on the model that was best supported, several nested models were tested about the invariance of the factors across grade levels. The nested models tested invariances in factor loadings, error variances, and factor correlations. Analyses were performed with LISREL 8.80 (Joreskog & Sorbom, 2006), based on the covariance matrices for each grade level.

Several goodness of fit indices were used to evaluate model results (Hoyle & Panter, 1995). The comparative fit index (CFI), nonnormed fit index (NNFI), and root mean square error of approximation (RMSEA) were used for overall model fit. The standardized root mean square residual (SRMR) and goodness of fit index (GFI) were used for individual subsamples. The χ^2, χ^2/df and the expected cross-validation index (ECVI) were used to compare overall and subsample models. Common rules of thumb were used in appraising the measures (Hoyle & Panter): .90 or more for CFI, NNFI, and GFI; .05 or less for RMSEA; .10 or less for SRMR; .05 alpha level for the χ^2, and 3 or less for χ^2/df.

Table 27 presents the overall fit indices for the three models. The fit indices for the one-factor solution were generally unsatisfactory, with the exception of the G4 subsample. The fit indices for the two-factor solution were generally satisfactory, with the exception of a low NNFI (.83) and high RMSEA (.14) for the overall analysis. The fit indices for the three-factor solution were generally satisfactory, with the same exceptions of a low NNFI (.87) and high RMSEA (.12) for the overall analysis. In summary, only the two- and three-factor solutions showed reasonable fit.

Comparisons of the two-factor and three-factor solutions showed that for both the overall and all subsample results, the χ^2 and χ^2/df differences were statistically (all *p*-values smaller than .001) and practically (all χ^2/df differences larger than 42) significant. In addition, the 90% confidence intervals for the ECVIs were .24–.26 and .15–.16 for the two- and three-factor solutions, respectively, indicating that the three-factor solution is expected to cross-validate better in a new sample. Therefore, the three-factor solution was chosen for further analysis. This solution is presented in Table 28.

Table 27

Tests of Invariance in Number of Factors

Model	*df*	χ^2	CFI	NNFI	RMSEA	SRMR	GFI
Three factors							
G4	11	523				.06	.96
G6	11	1,157				.07	.96
G8	11	1,651				.07	.96
G10	11	1,131				.06	.96
G12	11	1,052				.07	.96
Overall	55	5,514	.93	.87	.12		
Two factors							
G4	13	608				.07	.95
G6	13	1,414				.07	.95
G8	13	2,680				.08	.93
G10	13	2,013				.07	.93
G12	13	1,700				.08	.93
Overall	65	8,414	.90	.83	.14		
One factor							
G4	14	1,263				.10	.91
G6	14	3,209				.11	.88
G8	14	7,605				.14	.83
G10	14	8,245				.15	.80
G12	14	7,977				.16	.78
Overall	70	28,299	.65	.48	.23		

Note. CFI = comparative fit index; NNFI = nonnormed fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; GFI = goodness of fit index.

Table 28

Three-Factor Model: Fluency (F), Conventions (C), and Word Choice (W)

	G4	G6	G8	G10	G12
Factor loadings					
Essay length (F)	0.99	1.01	1.06	0.93	0.92
Style (F)	0.51	0.63	0.58	0.58	0.50
Grammar (C)	0.70	0.77	0.84	0.83	0.84
Usage (C)	0.63	0.62	0.61	0.61	0.59
Mechanics (C)	0.36	0.38	0.43	0.45	0.42
Vocabulary (W)	0.59	0.49	0.71	1.08	1.30
Word length (W)	0.60	0.68	0.79	0.81	0.85
Factor correlations					
F ↔ C	0.86	0.85	0.75	0.74	0.74
C ↔ W	-0.12	0.07	0.13	0.20	0.13
F ↔ W	-0.16	0.03	0.05	0.15	0.12
Error variances					
Essay length	0.09	0.01	0.02	0.07	0.07
Style	1.34	0.84	0.66	0.47	0.42
Grammar	0.32	0.37	0.31	0.37	0.33
Usage	0.29	0.48	0.69	0.75	0.71
Mechanics	0.82	0.79	0.79	0.86	0.87
Vocabulary	0.52	0.48	0.35	0.02	0.30
Word length	0.43	0.19	0.31	0.52	0.66

Table 29 presents fit indices for initial tests of invariance in the three-factor model. The three types of invariance, in factor loadings, factor correlations, and error variance, should be compared with the basic three-factor solution results in Table 27. These comparisons show that for all three invariances the overall χ^2 and χ^2/df differences are statistically (all p-values smaller than .001) and practically (all χ^2/df differences larger than 31) significant. In addition, the 90% confidence intervals for the ECVIs were .21–.23 and .24–.26 for the invariances in factor

correlation and error variance, respectively, indicating that the basic three-factor solution (without any of the invariance constraints) is expected to cross-validate better in a new sample than the solutions that presume invariances. For these two invariances, several subgroup results were also unsatisfactory.

Table 29

Tests of Invariance for Three-Factor Model

Invariance	*df*	χ^2	CFI	NNFI	RMSEA	SRMR	GFI
Factor loadings							
G4		863				.12	.94
G6		1,971				.12	.94
G8		1,811				.08	.96
G10		1,361				.10	.95
G12		1,797				.17	.91
Overall	83	7,804	.91	.88	.11		
Factor correlations							
G4		694				.09	.95
G6		1,240				.07	.95
G8		1,657				.07	.96
G10		1,199				.08	.96
G12		1,092				.08	.96
Overall	67	5,882	.93	.89	.11		
Error variances							
G4		2,091				.15	.83
G6		1,507				.08	.94
G8		1,755				.07	.96
G10		1,623				.07	.95
G12		1,648				.08	.95
Overall	83	8,625	.90	.87	.12		

Note. CFI = comparative fit index, NNFI = nonnormed fit index, RMSEA = root mean square error of approximation, SRMR = standardized root mean square residual, GFI = goodness of fit index.

However, for the factor correlations invariance, the 90% confidence intervals for the ECVI were .16–.17, similar to that of the basic three-factor solution (.15–.16). In addition, the fit of this nested model was slightly better than the basic three-factor model, with an NNFI of .89 and an RMSEA of .11. Therefore, the three-factor model with invariant factor correlations may be considered satisfactory. The three invariant correlations are .77 between fluency and conventions, .12 between conventions and word choice, and .07 between fluency and word choice.

Discussion

The goal of this project was to develop a tool for measuring essay writing performance across grade levels in a way that allows within-grade normative comparisons and cross-grade developmental comparisons. Based on objective measures that are related to human-rated writing performance measures from early elementary school up to post-secondary levels, a single unified score scale was constructed.

Summary of Results

The scale was constructed on the basis of a large national sample of 170 schools and about 12,000 students in Grades 4, 6, 8, 10, and 12. Analyses suggested that the school sample was reasonably representative of the population. Moreover, correcting discrepancies between the student sample and the population in terms of school characteristics (school type, locality, region, minority enrollment, and eligibility for free or reduced-priced lunches) had almost no effect on grade-level performance distributions.

Standardized scale scores were based on the weighted average of seven *e-rater* features, where the weights were derived from a factor analysis of the features. Differences in grade-level performance were larger for lower grade levels (average scale scores of -1.06 for 4th grade and -0.39 for 6th grade) than for higher grade levels (0.51 for 10th grade and 0.75 for 12th grade). The within-grade standard deviations were slightly smaller for 4th grade (0.73) than for higher grade levels (0.76–0.82). Grade-level score distributions also showed increasing (negative) skewness and (positive) kurtosis with grade level, reflecting an increasing lower tail in performance of higher grade levels.

A cross-classified random-effects model estimated the variance that could be attributed to students and topics and investigated the importance of different essay, student, and topic characteristics on scale scores. The feasibility of a single cross-topic score scale was supported

by the small variance component for topics—topics accounted for less than 2% of the score variance (when grade level was controlled), and half of this amount was explained by mode of writing. In turn, the features responsible for the mode differences (lower scores for persuasive essays) were the two word-level features (vocabulary and word length).

Among the findings related to student characteristics are the higher scores for females (by 0.31 on average) and higher scores for white students (0.10). Surprisingly, ESL students did not earn lower scores on average than non-ESL students (conditional on all other variables in the model). English schooling was the major linguistic background factor that influenced performance, with an average increase in scores of 0.16 from students who attended English schools less than a year, to those who attended English schools more than a year, to those who always attended English language schools.

The measurement of student writing performance using the scale scores is quite reliable. The median grade-level reliability estimate for a single essay is .74 (within mode). The high reliability of the scale scores, together with previous results on the near-unity true-score correlation between *e-rater* scores and human scores (Attali & Burstein, 2006), suggests that the scale scores provide adequate measurement of essay writing ability.

The feasibility of the unified developmental scale was also supported in a longitudinal study. Increase in performance of 401 repeating students across a 1-year interval conformed to the expected increase according to the main cross-sectional study.

The feasibility of the unified developmental scale was further supported in a human scoring experiment. Human raters scored essays that they thought were written by either 6th or 10th graders. In both cases, the essays were actually written by students from Grades 6, 8, and 10. The average human scores for students across these three grade levels matched their average automated scale scores. These results support the assumption of a unified scale that scale scores are equally as sensitive as human raters to developmental differences in writing quality across grade levels.

Finally, exploratory and confirmatory factor analysis of the seven features revealed a similar underlying structure across grade levels. The three-factor structure that was best supported by the data can be conveniently described as a fluency (with the essay length and style features), conventions (with grammar, usage, and mechanics), and word choice (vocabulary and word length) structure. The confirmatory factor analysis found reasonable support for this

structure across all five grade levels. Furthermore, some support was found for a constrained three-factor solution with the same (invariant) factor correlations across grade levels. The correlations are .77 for fluency and conventions, .12 between word choice and conventions, and .07 between word choice and fluency.

Possible Applications

One possible use of the developmental writing scale is as a yardstick for interpreting performance levels in different assessment programs. This may be particularly useful for state assessments that could interpret their results in relation to national developmental norms. A related application would provide context for interpreting trends in performance of assessment results.

Another application of the developmental scale would be to provide enhanced interpretation of individual essay scores. Instead of the typical holistic 1–6 scale, scores based on the developmental scale provide the user (student, teacher, or parent) with a richer normative interpretation of performance. For example, in *Criterion*, ETS's writing instruction application, students from Grade 4 until college receive automated instructional feedback and essay scores on their writing. Essay scoring is performed by *e-rater* scoring models for each grade level. These models are based on samples of essays scored by human raters. They could be replaced by scores based on the developmental scale, thus providing a more unified and consistent scoring approach.

The data upon which the developmental writing scale is based can also be used to investigate new potential automated measures of writing quality. The developmental nature of the data provides an alternative criterion for evaluating potential measures, one that is not based on human holistic ratings of essays. By analyzing the distribution of the potential measure both within and across grade levels, one can develop a better sense of the measure's suitability for AES. For example, sentence length is an important measure of syntactic complexity in reading research and is the only measure for syntactic complexity in the Lexile Framework for Reading (www.lexile.com), which provides tools for assessing difficulty of reading texts. However, when this measure is analyzed in the present context of essays written by children, a negative correlation between sentence length and grade level is revealed. That is, average sentence length in an essay is generally shorter for older children, which may be due to better punctuation and sentence structure for older children.

In this context, the fluency, conventions, and word choice structure can also guide further development of automated measures of writing quality. These dimensions of writing seem to be well suited for a theory of writing for novices, such as the knowledge telling model of writing (Bereiter & Scardamalia, 1987). It may also be more suitable for AES technology.

Limitations

Several limitations of the scale should be noted. The first is that the scale is based on essays written with a particular time limit: 30 minutes. Previous research (reviewed in Breland et al., 1999) showed that essay scores significantly increase by allowing more time for writing. This finding limits the applicability of the scale to essays written under different time limits. However, there is no research on the causes of score increases with longer time limits. It is reasonable to assume that the aspects of writing that would benefit most from longer time limits are those associated with discourse and fluency and not with word and sentence level aspects of writing. In future development of automated scoring it might be possible to standardize the measurement of these higher-level aspects of writing with respect to amount of time allowed for the student, similarly to the way the grammar score in *e-rater* takes into account the length of the essay in weighting grammar errors.

A second limitation of the scale is that it is based on essays written in two modes of writing: descriptive and persuasive. Although the differences in scores between the two different modes were small, and analyses showed that these differences result mainly from the word choice features, more research on other modes of writing is clearly needed in order to incorporate them into the developmental writing scale.

A final limitation of the scale is that it is based only on the specific features used in this study and thus is limited by what these features measure. Even though past research showed very high correlations between *e-rater* scores and human essay scores, it is clear that important dimensions of writing are not represented in the feature set.

References

Attali, Y. (2006, April). *On-the-fly automated essay scoring*. Paper presented at the annual meeting of the National Counsel of Measurement in Education, San Francisco, CA.

Attali, Y. (2007). *Construct validity of* e-rater *in scoring TOEFL essays* (ETS Research Rep. RR-07-21). Princeton, NJ: ETS.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with *e-rater* V.2. *Journal of Technology, Learning, and Assessment, 4*(3). Retrieved March 21, 2008, from http://escholarship.bc.edu/jtla/vol4/3/

Ben-Simon, A., & Bennett, R. E. (2006, April). *Toward theoretically meaningful automated essay scoring*. Paper presented at the annual meeting of the National Council of Measurement in Education, San Francisco, CA.

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Breland, H. M., Bridgeman, B., & Fowles, M. E. (1999). *Writing assessment in admission to higher education: Review and framework* (College Board Report No. 99-3; GRE Board Research Report No. 96-12R). New York: College Entrance Examination Board.

Breland, H. M., Camp, R., Jones, R. J., Moris, M. M., & Rock, D. A. (1987). *Assessing writing skill* (Research Monograph No. 11). New York: College Entrance Examination Board.

Burstein, J. C., Kukich, K.,Wolff, S., Lu, C., & Chodorow, M. (1988, April). *Computer analysis of essays*. Paper presented at the annual meeting of the National Council of Measurement in Education, San Diego, CA.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum

Elliot, S. M. (2001, April). *IntelliMetric: From here to validity*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Gruber, K. J., Wiley, S. D., Broughman, S. P., Strizek, G. A., & Burian-Fitzgerald, M. (2002, May). *Schools and staffing survey, 1999-2000: Overview of the data for public, private, public charter, and Bureau of Indian Affairs elementary and secondary schools* (NCES 2002–313). Retrieved March 21, 2008, from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2002313

Hoyle, R. H., & Panter, A. T. (1995). Writing about structural equation models. In R. H. Hoyle (Ed.), *Structural equation modeling—Concepts, issues, and applications* (pp. 158–176). Thousand Oaks, CA: Sage.

Joreskog, K., & Sorbom, D. (2006). LISREL 8.80 [Computer software]. Chicago: Scientific Software International.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes, 25,* 259–284.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15–21.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan, 48*, 238–243.

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

Persky, H. R., Daane, M. C., & Jin, Y. (2003, July). *The nation's report card: Writing 2002* (NCES 2003–529). Retrieved March 21, 2008, from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2003529

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). New York: Macmillan.

National Center for Education Statistics. (2006a). *2001–2002 private school universe survey data.* Retrieved March 21, 2008, from http://nces.ed.gov/surveys/pss/privateschoolsearch/

National Center for Education Statistics. (2006b). *2003–2003 public elementary/secondary school universe survey data.* Retrieved March 21, 2008, from http://nces.ed.gov/ccd/pubschuniv.asp

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Sage Publications.

Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2004). HLM6: Hierarchical linear and nonlinear modeling [Computer software]. Chicago: Scientific Software International.

Resnick, L. B., & Resnick, D. P. (1990). Tests as standards of achievement in schools. In G. R. Anrig (Ed.), *The uses of standardized tests in American education: Proceedings of the 1989 ETS invitational conference* (pp. 63–80). Princeton, NJ: ETS.

Notes

[1] *Criterion* is a writing instruction application developed by ETS, where students receive automated feedback and essay scores (based on *e-rater*) on their writing.

Appendix
School Sampling Plan

In this appendix, the sampling plan for schools invited to the study is reviewed. Changes made in the second year of the study are noted and explained.

The overall plan was to recruit a national sample of schools that would contribute at least one class for each grade that was included in the study (4th, 6th, 8th, 10th, and 12th grades) and present in the school. These classes wrote two essays in each of two writing modes during the spring of 2005 (and the study was repeated in 2006). The topics for each class were randomly selected from a pool of topics, such that each topic could be assigned to classes of two or three consecutive participating grades (for example, a sixth-grade topic was presented also to fourth graders and eighth graders).

Databases for School Sampling

The purpose of the school sampling plan was to invite a national sample of schools to participate in the study by registering classes in the grade levels included in the study. The sampling of invited schools was based on information extracted from the latest available (in October 2004) comprehensive national school surveys prepared by the NCES. The public school information was extracted from the *2002–2003 Public Elementary/Secondary School Universe Survey Data* (NCES, 2006b). The private school information was extracted from the *2001–2002 Private School Universe Survey Data* (NCES, 2006a).

School background information included in these two databases was used to ensure the representativeness of the invited schools sample. However, because it was assumed that only a small number of invited schools would be interested in participating, the school background information was also used to assess the representativeness of the participating schools sample and to correct for any discrepancies between the sample and the population of schools. In addition, after the first year, an interim analysis of the participating sample was used to try to increase the representativeness of the sample by oversampling certain types of invited schools in the second year.

Preliminary Considerations in School Sampling

Although the plan was to test classes in schools, invitations had to be sent to schools. The invitation letter encouraged schools to include at least one class per participating grade. However, sampling of invited schools was based (among other criteria) on the number of students in relevant grades in the school (this information was available in both databases). In other words, it was assumed that larger schools would contribute more students to the study.

On the basis of the Schools and Staffing Survey of the NCES, Gruber, Wiley, Broughman, Strizek, & Burian-Fitzgerald (2002) report that the average public school class size is 4% larger than average private school classes in lower grades and 15% larger in higher grades. A rough adjustment in the initial sample plan was to sample 10% more private schools than their relative share.

The sampling for invited schools used both explicit and implicit stratification of important variables to assure adequate coverage of the national population of schools. These will be described in the next sections.

Public School Invitation Sampling Plan

There were 99,635 records in the public school database. School records were retained based on the following criteria: 95% of the total school records were operational; 91% of the operational schools were regular (and not special education, vocational, or other/alternative schools); 96% of the schools retained were defined as elementary, middle, or high schools (and not *other*); 98% of the schools retained were located at one of the four regions defined below; 100% of those were located at one of the location types defined below; 94% of those had complete data with respect to the calculation of percent of free/reduced priced lunch students and minority students; and 94% of those had students in participating grades (excluded were elementary schools up to third grade or middle schools with only fifth grade). The final number of school records was 72,403, or 73% of the initial number of records. Table A1 summarized the number of schools and students in this final database.

The explicit stratification of public schools was based on three school-level variables that significantly contribute to variance of NAEP writing scores (Persky, Daane, & Jin, 2003). The first variable was school region, with four regions defined according to the NAEP definition (northeast, southeast, central, and west) except for Virginia schools, which were all included in the southeast region. The second variable was location. The original information provided in the database

included eight types of school locations, which were collapsed into three types: city (including large- or midsize city), urban fringe (including large towns), and rural (including small towns). The third variable was based on the percent of students in the school that are eligible for free or reduced-priced lunches. The 33.3 and 66.7 percentile ranks of these values were computed separately for each school level and each school was categorized to one of three levels of free/reduced-price lunch student eligibility. These three explicit stratification variables form a total of 36 strata (4 x 3 x 3 levels).

Table A1

Number of Public Schools and Students, by Grade and School Level

	Elementary		Middle		High	
	Schools	Students	Schools	Students	Schools	Students
4	44,051	3,217,061	1,181	148,460	0	0
6	17,933	1,011,488	11,150	2,397,618	0	0
8	4,498	202,429	12,940	2,911,974	2,881	161,805
10	0	0	0	0	13,792	3,059,683
12	0	0	0	0	13,824	2,570,019
Total	44,078	4,430,978	14,488	5,458,052	13,837	5,791,507

Within each stratum, schools were sorted according to one implicit stratification variable based on the percent of minority students in the school (Hispanics, Black, Asian, or Native American). The 33.3 and 66.7 percentile ranks of these values were computed separately for each school level, and each school was categorized to one of three levels of minority enrollment.

Finally, each school was sampled with probability proportional to the number of students that were enrolled in the grades relevant to the study. This was accomplished by first creating a grand index of relevant students at each of the 36 strata cells, where the relevant students in the first school of the cell occupy positions 1 to the number of students in that school, and students of following schools occupy subsequent positions. Calculations were made to find the total number of students in relevant grades (T), the proportion of students in each of the 36 cells (P), and the total number of schools that should be invited (S, see below for the number of schools invited). The number of schools selected in each of the 36 cells should be proportional to the

number of relevant students in this cell. This number, N, is equal to S multiplied by P (the proportion of students in that cell), rounded to the nearest whole number. To find what schools to invite, a random number was generated between 1 and the T/N to provide the initial step on the grand index. Beginning with this initial step, the grand index was stepped N times in intervals of T/N. The positions that were stepped on are associated with the schools that were invited.

Private School Invitation Sampling Plan

There were 23,127 private schools with 1,751,645 students in relevant grades in the private school database. Of those, 18,300 schools (79% of schools) with 1,552,404 students (89% of students) were defined as regular elementary or secondary schools (and not Montessori, special program emphasis, special education, vocational/technical, early childhood program, or alternative).

The explicit stratification of private schools was based on the region and location of the school (with a total of 12 strata), and within each stratum schools were sorted according to the minority level (three levels based on the 33.3 and 66.7 percentile ranks of these values in private schools). Finally, each school was sampled with probability proportional to the number of students that were enrolled in the grades relevant to the study, similar to the public school invitation plan.