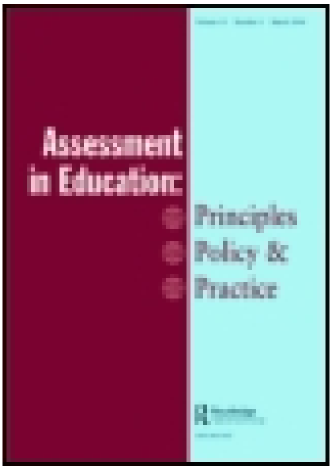

Assessment in Education: Principles, Policy & Practice
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/caie20

How important is content in the ratings of essay assessments?

Mark D. Shermis [a], Aleksandr Shneyderman [b] & Yigal Attali [c]
[a] Department of Educational Psychology, University of Florida, Gainesville, FL, USA
[b] Department of Assessment, Research, and Data Analysis, Miami-Dade County Public Schools, Miami, FL, USA
[c] Educational Testing Service, Princeton, NJ, USA

Published online: 13 Feb 2008.

To cite this article: Mark D. Shermis, Aleksandr Shneyderman & Yigal Attali (2008) How important is content in the ratings of essay assessments?, Assessment in Education: Principles, Policy & Practice, 15:1, 91-105, DOI: 10.1080/09695940701876219

To link to this article: http://dx.doi.org/10.1080/09695940701876219

Assessment in Education: Principles, Policy & Practice
Vol. 15, No. 1, March 2008, 91–105

How important is content in the ratings of essay assessments?

Mark D. Shermis*[a], Aleksandr Shneyderman[b] and Yigal Attali[c]

[a]*Department of Educational Psychology, University of Florida, Gainesville, FL, USA;* [b]*Department of Assessment, Research, and Data Analysis, Miami-Dade County Public Schools, Miami, FL, USA;* [c]*Educational Testing Service, Princeton, NJ, USA*

This study was designed to examine the extent to which 'content' accounts for variance in scores assigned in automated essay scoring protocols. Specifically it was hypothesised that certain writing genre would emphasise content more than others. Data were drawn from 1668 essays calibrated at two grade levels (6 and 8) using *e-rater*™, an automated essay scoring engine with established validity and reliability. *E-rater* v 2.0's scoring algorithm divides 12 variables into 'content' (scores assigned to essays with similar vocabulary; similarity of vocabulary to essays with the highest scores) and 'non-content' (grammar, usage, mechanics, style, and discourse structure) related components. The essays were classified by genre: persuasive, expository, and descriptive. The analysis showed that there were significant main effects due to grade, $F(1,1653) = 58.71$, $p < .001$, and genre $F(2, 1653) = 20.57$, $p < .001$. The interaction of grade and genre was not significant. Eighth grade students had significantly higher mean scores than sixth grade students and descriptive essays were rated significantly higher than those classified as persuasive or expository. Prompts elicited 'content' according to expectations with lowest proportion of content variance in persuasive essays, followed by expository and then descriptive. Content accounted for approximately 0–6% of the overall variance when all predictor variables were used. It accounted for approximately 35–58% of the overall variance when 'content' variables alone were used in the prediction equation.

Introduction

This study was designed to examine the extent to which content accounts for variance in scores assigned in automated essay scoring protocols. In the context of this study, content refers to the degree to which a writer addresses the topic of a prompt. If the prompt were a question, the content of the response would be an assessment of how well the writer brought to bear relevant facts and information to answer the question (i.e., the correctness of the answer). Leki and Carson (1997) refer to this as 'text-responsible' prose since the response is based on ostensibly identifiable sources of information. Shermis et al. (2001) found a relation between the type of content elicited by prompts and the way in which raters evaluated essays. The process goes something like this: prompts elicit themes that can be reliably classified, the classification is taken into consideration by the rater, and the rater enters this consideration into the formulation of a score. Based on earlier work by McTavish et al. (1997) and Rajecki et al. (1993), content was classified according to one of four general themes: traditional, practical, emotional, and analytic, as employed in the *Minnesota Contextual Content Analysis* (*MCCA*) program.

With the traditional theme, normative considerations are central. Matters are viewed in terms of standards, rules and codes that guide social behaviour. This theme is found in judicial and religious circles. The focus of the practical theme is on pragmatism; that is, successful and practical

*Corresponding author. Email: mshermis@ufl.edu

ISSN 0969-594X print/ISSN 1465-329X online
© 2008 Taylor & Francis
DOI: 10.1080/09695940701876219
http://www.informaworld.com

goal accomplishment. This theme is found in business and work dealings. With the emotional theme, affective perspective predominates; personal involvement, comfort, and enjoyment are highlighted. This theme is found in the areas of close personal relationships, leisure, and recreation. A cognitive perspective predominates in the analytic theme and the situation or activity is defined in terms of objectivity, curiosity, or interest. This theme is found in research and educational settings (McTavish et al. 1997).

Background

An interesting finding in that earlier study was that themes elicited by practical or analytic prompts tended to be rewarded more highly (assigned higher scores) by raters than themes linked to traditional or emotional prompts. Psychometrically, when evaluation of content becomes a part of the scoring process, the type of elicited content effectively influences the empirically-determined difficulty level of the essay. Additionally, the fact that prompts elicit specific themes leads to several important implications for assessment practice, including (a) the refinement of procedures for developing rubrics based on targeted outcomes; (b) the reduction of rater bias in subjective ratings; and (c) the understanding of cognitive bases for categorisation employed by raters.

The Shermis et al. (2001) study in part replicated an early undertaking by Hamp-Lyons and Mathias (1994) who used expert judges to evaluate the difficulty level of ESL prompts. These researchers classified prompts along two axes: public–private and expository–argumentative. A ‘public’ orientation reflects the situation where a writer is expected to speak about groups and communities in contrast to describing themselves and/or their families. A ‘private’ orientation means that writers are asked to say how they feel about something or to use personal experience in their response. Expository writing refers to factual writing and includes the genres of narrative and descriptive essays. Argumentative writing addresses the situation where the writer takes a position and tries to justify it and/or persuade others to share it (i.e., persuasive essays) (Hamp-Lyons and Mathias 1994).

It was hypothesised that the most difficult prompts would be argumentative-public followed by the easier categories of expository–public, argumentative–private, and expository–private. The difficulty levels, as judged by the expert raters, were consistent with the hypothesised rank-ordering, but the empirically-derived difficulty levels were *inversely* related. The researchers speculated that the discrepancy between how raters view essays and how they score them might be attributable to some compensating mechanism. That is, raters may be more generous to even marginal responses on prompts that they view as being ‘hard’ than they would to essays they perceive of as being ‘easy’.

Both the Hamp-Lyons and Mathias (1994) and Shermis et al. (2001) are a subset of studies that have sought to determine what factors influence prompt difficulty. Many things could influence how difficult or easy a topic might be perceived or evaluated as being. For instance, asking students to write an essay one hour before being dismissed for winter break could conceivably impact on their normal writing performance. In a similar vein, asking raters to evaluate essays without the benefit of training might influence their judgements, and consequently, the calibration of prompt difficulty levels.

Ruth and Murphy (1988) have summarised the most important research on systematic factors that influence writing difficulty. These include (1) knowledge requirement of subjects (Are the writers familiar with the subject matter/content or is it foreign/unique to them?); (2) student by subject interaction (Are writers allowed to choose the topic from a list of possible essays or are they all assigned the same writing task?); (3) wording effects (Are writers generating essays to a specific audience or purpose or is the task more general?); and (4) bias (these typically are unknown cultural factors that might influence writing performance, e.g., writing on a topic that is personally or politically sensitive). Other aspects of the writing assignment can also influence prompt difficulty as well including task instructions, the type of discourse, and the testing context.

Rationale

The present study focused on the identification of the amount of overall variance accounted for by content. To accomplish this objective, prompts were linked to specific writing genres, and then the relation of the scoring to the output was mapped. Our analysis of demands made by prompts suggested that the emergent outputs of prompts are scored differently according to raters' estimates of relevance and importance of the content classification. This point can be illustrated by examining the demands made by each of three prompts:

Example 1. Provide a narrative of Christopher Columbus's discovery of America including key dates, individuals, and events.

Example 2. Argue for or against the following proposition: Volunteerism should be an integral component of undergraduate education.

Example 3. What is the most important problem facing American society today? Describe the nature of the problem and why you think it is significant.

The first prompt demands a display of content for achieving a high score. Raters will presumably look for enumeration of facts, descriptions of sequences, and identification of characters. Keywords, synonyms, word nets, or other appropriate references provide the bases for determining the relevance of the content in addressing the prompt demands. On the other hand, content is only one component of the second prompt. The successful respondent would be familiar with the ideas of volunteerism and the undergraduate curriculum, but more critically the successful product depends on the respondent's rhetorical skills for persuasive argumentation. The last prompt is characteristic of what is referred to in psychology of thinking as an 'ill-structured' problem, i.e., problems without an obvious path to the solution as would be characteristic of high school physics or geometry problems. Roughly speaking, the prompt demands selection of relevant ideas from a knowledge base, integration of those ideas around the problem statement, insightful reasoning to evaluate potential solutions and, eventually, arriving at an acceptable solution. As can be seen, this prompt has its focus on process; it is secondarily linked to content, thereby depending more on general problem-solving or critical thinking skills than on content.

There is no universally agreed-upon list of writing genres, but of the many proposed, four long-standing categories emerge: description, narration, exposition, argument (D'Angelo, 1976).[1] The purpose of description is for the writer to convey a sense experience while narration is a category in which a writer tells a story or relates an event. The exposition category is designed for the writer to present ideas whereas argument (i.e., persuasive writing) is a type of discourse where the writer convinces others to a particular point of view. D'Angelo (1976) provided a taxonomy for each genre which is re-produced in Table 1. Shermis et al. (2005) examined the genre labels in all US statewide elementary and secondary-level high- stakes testing programmes and found the four labels subsume the genres (i.e., sometimes under slightly different names) covered under the testing programmes.

Attali (2006) explored the issue of using genre as a prompt difficulty factor in a larger study in which he attempted to establish norms for automated essay scoring. In that study he used both grade level and genre to statistically model ratings assigned by human raters, and then tested the model to determine how well the computer replicated the human assessments. He found that both factors worked well in establishing generic norms that could be applied to a variety of prompts within each grade level and genre. The study was the first attempt to create a set of norms in which comparisons could be made across prompts with different difficulty levels. Prior to his work,

Table 1. Bain's forms of disclosure (derived from D'Angelo, 1976).

	Function	Subject	Organisation	Language
Description	Evoke sense experience	Objects of senses	Space/time	Denotative and connotative, figurative, literal, impressionistic, objective
Narration	Tell a story, narrate an event	People and events	Space/time	As above
Exposition	Inform, instruct, present ideas	Ideas, generalisations	Logical analysis and classification	Denotative and factual
Argument	Convince, persuade, defend, refute	Issues	Deduction and induction	Factual and emotive, depending on appeal

Source: Ruth and Murphy (1988, 89).

calibrations of difficulty were made at the prompt level, so comparisons from one prompt to the next were confounded by differing sets of norms.

Automated Essay Scoring (AES) is the evaluation of written work via computers. Initial research restricted AES to English, but has recently extended to other languages as well (Vantage Learning 2001, 2002; Kawate-Mierzejewska 2003). Most packages place documents within an electronic portfolio. They provide a holistic assessment of the writing which can be supplemented by trait scores based on an established rubric, and may provide qualitative critiques through discourse analysis. Most use ratings from humans as the criterion for determining accuracy of performance, though some of the packages will permit validation against other sources of information (e.g., large informational databases).

Obviously, computers do not 'understand' written prose in the same way that humans do, a point that may be unnerving until one reflects on ways alternative technologies achieve similar results. Thus, one can estimate the length of a wall using a traditional tape measure or employ a laser-pointing device to achieve similar results. The computer scores essays according to models of what human raters consider desirable and undesirable writing elements. Collections of these elements are referred to as 'traits', the intrinsic characteristics of writing called 'trins' (Page and Peterson 1995). The specific elements are called proxies or 'proxes' (Page and Petersen 1995). The differentiation of 'trins' and 'proxes' is parallel to that of 'latent' and 'observed' variables in the social sciences: thus, the score on an IQ test might be thought of as a 'prox' (specific element) for the underlying characteristics of the 'trin' (conceptualisation) intelligence.

AES software packages include computer programs that parse the essay text, for the purpose of identifying hundreds of prox variables ranging from simple to complex. A deceptively simple variable is essay length. Although raters value this attribute, the relationship to good writing is not linear but rather logarithmic; i.e., raters value the amount of writing output up to a point, but then they look for other salient aspects of writing once the quantity threshold is met. Similarly, the number of occurrences of 'because' is a relevant feature. Although seemingly a superficial feature, it importantly serves as a proxy for the beginning of a dependent clause. And this, in turn, is reflective of sentence complexity.

When human raters comprise the criterion against which rating performance is judged, AES engines work off a statistical model developed using the following procedures: (1) obtain a sample of (500) essays with (4–8) human ratings on each essay; (2) randomly select (300) essays and regress the human ratings against the variable set available from various computational analyses of a text; (3) use a subset of consolidated feature variables, or the factor structure underlying a set of feature variables, in order to formulate a regression equation. The equation does not have to have a linear basis, but linear models are easier to explain; (4) cross-validate the regression equation on the (200) remaining essays to determine if the original regression line has suffered from shrinkage (Shermis et al. 2006).

Most of the evidence suggests that AES evaluations are equivalent to or higher than evaluations of reliability with human raters (Elliot 2003; Landauer et al. 2003). All AES engines have obtained exact agreements with humans in the mid 80s and adjacent agreements in the mid to high 90s – slightly higher than the agreement coefficients for trained human raters. The slight edge for AES may be a function of the fact that the statistical models are based on more raters than one would typically find in a rating enterprise. Several validity studies have suggested that AES engines tap the same construct as that being evaluated by human raters. Page et al. (1995) examined the construct validity of AES, Keith (2003) summarised several discriminant and true score validity studies of the technology, and Attali and Burstein (2006) demonstrated the relationship between AES and instructional activities associated with writing.

AES is not without its detractors. Ericcson and Haswell (2006) performed a comprehensive critique of the technology from the perspective of those who teach post-secondary writing. Objections to the technology ranged from a concern about the ethics of using computers rather than humans to teach writing to the lack of synchronicity between how human graders approach the rating task and the process by which AES evaluates a writing sample, to failed implementations of AES in university placement testing programmes. Nevertheless, AES is an increasingly pervasive assessment technology that is used for both assessment and instruction.

Because of the way in which AES is formulated, it incorporates any rater-prompt interactions that might be present with human raters. Engelhard (1992) used a multifaceted Rasch measurement model to investigate the rater severity and other aspects of writing competence assessment in the high-stakes Eighth Grade Writing Test administered annually in the state of Georgia. Each of the 1000 randomly selected essays were graded by 2 out of a total of 82 raters. The raters used analytic/primary trait mode of scoring, in which five domains of writing (content/organisation, style, sentence formation, usage, and mechanics) were rated. The author investigated four facets of writing assessment: students' writing ability, the difficulty of the assigned writing task (prompt), the difficulty of the writing domain, and the severity of the rater. He concluded that although the prompts showed significant differences in difficulty levels, these differences were small, and not practically important. The sentence formation domain was found to be significantly less difficult than any other domain. As for rater severity, the author found that only about 20 of the 82 raters could be considered as having approximately equal 'middling' severity. Severities of the remaining raters varied substantially from the low of −1.74 for the most lenient to the high of +1.78 for the most severe rater (on the logit scale). Engelhard pointed out that these differences in rater severity were exhibited despite the fact that the raters were highly trained.

Wolfe (2004) used the multifaceted Rasch model to identify rater effects. In his formulation, only student writing competence and rater severity were used as facets of measurement; the prompts were assumed to have equal difficulty. The author identified the following rater effects: accuracy/inaccuracy, severity/leniency, and centrality/extremism; he pointed out that '…each of these effects exists in the data in both non-ignorable rates and magnitudes.' (Wolfe 2004, 35).

One of the more comprehensive studies of rater-prompt interactions was conducted by Weigle (1998) using ESL essays. Her study combined both quantitative (multi-faceted Rasch) and qualitative components (read aloud protocols) and was directed at evaluating the impact of rater experience. The results were mixed: on one prompt, inexperienced raters were more severe than experienced raters, but not on the other prompt. The results of the qualitative analysis suggested that differences may have been related to the ease with which the scoring rubric was applied to prompt grading, and to the raters' perceived appropriateness of the prompts.

The findings by Attali (2006) regarding the independent ideas of prompt demands and the classification of genre led to the hypothesis that scores assigned to essays are dependent, at least in part, on the raters' concern with content importance as it is constrained by explicit writing genres. Specifically, content was assumed to play an increasing role as one proceeds through the genre series from persuasion and narration to expository writing products. Effectively mapped, the genre classifications have the potential for production of prompts to guide effective and efficient automated essay scoring calibrations. Thus, if one wished to design a 'persuasive' prompt, and it was known that content played only a marginal role in the evaluation of the product, then the general model of writing would be based on genre characteristics rather than on content features. This could have a significant impact on sample sizes required for prompt development and calibration, and could make it easier for teachers to create their own prompts 'on the fly'.

In summary, the study addresses the issue of whether or not content plays a major role in determining essay scores.

Method

Sample

The data were drawn from N = 7575 essays contained in one of the standardisation samples in the ETS *Criterion*^{SM} database. (*Criterion* is the instructional and portfolio component of a system that incorporates *e-rater*™ as an automated essay scoring component. *Criterion* and *e-rater* are described in more detail below.) These 7575 essays were written by students in grades 6–12. Essays were solicited from a cluster of 480 K–12 clients (districts or schools) across the United States comprising a pool of 160,000 users, and were collected under 'practice-test' conditions. No demographic data were gathered at the student level. Prompts were distributed across grade levels as shown in Table 2.

Table 3 shows the distribution of the 36 prompt responses by genre and grade levels.

The original ETS prompt taxonomy was more specific than the hypothesised genre (i.e., description, narration, exposition, persuasion) and was not replicated across all grades. An attempt was made to reclassify the original ETS genre with the proposed scheme. For example, in grade 6, the nature of the 'How To' prompts was determined to be sufficiently close to exposition prompts to warrant a reclassification. Alternatively, the 'Problem and Solution' prompts did not match well any of the proposed genre. After reclassification, the essays were consolidated into a sampling matrix displayed in Table 4. A decrease in the number of essays for most grade levels compared to the previous classification reflects the fact that certain original prompts could not be classified within the proposed genre taxonomy. This prompt reclassification reduced the number of essays in the sample from N=7575 to N=5869. Essays of students in grades 6 and 8 were selected for subsequent comparisons because none of the other grade levels reflected all three proposed genre: descriptive, expository, or persuasive. Note that after sampling and re-classification there was no 'narration' genre, and only 578 of the 1017 6th-graders could be used for comparison purposes. Thus, the number of essays was reduced once more from N = 5869 to N = 1668 (i.e., 576 6th-grade essays and 1092 8th-grade essays). Of those, nine essays were not scored by the *e-rater* because the values of one of the *e-rater* features were missing. These essays were not used in the analyses involving essay scores. The number of prompts was reduced from the original 36 to 8 (three 6th-grade prompts and five 8th-grade prompts.)

Table 2. Distribution of criterion prompts across grades 6–12.

Grade level	6	7	8	9	10	11	12	Total
Number of prompts	5	4	5	4	7	6	5	36

Table 3. The distribution of essays across genre and grade.

	Grade							
Genre	6	7	8	9	10	11	12	Grand total
Cause-and-effect	220					181	145	546
Comparison and contrast		212	219	152	206	217		1006
Descriptive	220	219	219					658
How-to	140		219					359
Persuasive	218	219		219	220	218	217	1311
Problem and solution					219	223		442
Response to literature	219			220	220		219	878
Workplace writing					218			218
Writing for assessment		199	435	219	437	433	434	2157
Grand total	1017	849	1092	810	1520	1272	1015	7575

Instruments

Criterion^{SM} is a web-based service developed by ETS for evaluating writing skills, instantaneous reports of scores and diagnostic feedback. (For a detailed description of the system, see Burstein et al. (2003). *Criterion* incorporates two complementary applications based on Natural Language Processing methods. One application, *e-rater,* extracts linguistically-based features from an essay and uses a statistical model to determine how these features are related to overall writing quality, so that a holistic score may be assigned to the essay. The second application, *Critique*, is comprised of a suite of programs that evaluates and provides feedback for errors in grammar, usage, and mechanics, identifies the essay's discourse structure, and recognises undesirable stylistic features.

The writing analysis tools in *Critique* are used to identify five main types of grammar usage, and mechanical errors including agreement errors, verb formation errors, wrong word use, missing punctuation, and typographical errors. The detection of grammatical violations is corpus based and statistical.

The construction of *e-rater* v2.0 models is given in detail in Attali and Burstein (2006). It is composed of 12 features used by *e-rater* v2.0[2] to score essays. The 12 features are associated with six areas of analysis: errors in grammar, usage, and mechanics (Leacock and Chodorow 2003); style (Burstein 2003); identification of organisational segments, such as thesis statement (Burstein et al. 2003); and vocabulary content (Attali and Burstein 2006).

Eleven of the individual features reflect essential characteristics in essay writing and are aligned with human scoring criteria. The first 6 of the 11 features are contained in the *Critique* writing

Table 4. The distribution of essays across the consolidated genre and grade.

	Grade							
Genre	6	7	8	9	10	11	12	Grand total
Descriptive	220	219	219					658
Expository	140		438	152	220	218	145	1313
Persuasive	216	418	435	438	1082	656	651	3898
Grand total	576	637	1092	590	1302	874	796	5869

analysis tools, and reflect the kinds of feedback that human raters provide, though not necessarily in the same statistical form (Attali 2004). These features include: (1) squared proportion of errors in grammar, (2) proportion of word usage errors, (3) proportion of mechanical errors, (4) proportion of style comments, (5) number of required discourse elements, (6) average length of discourse elements, (7) score assigned to essays with similar vocabulary, (8) similarity of vocabulary to essays with score ‘6’, (9) number of word types divided by number of word tokens, (10) log frequency of least common words, (11) average length of words, and (12) total number of words (Attali and Burstein 2006). Features 7 and 8 in this study are referred to as ‘content-related’ features.

Once the values of all 12 features are determined, *e-rater* uses them to score essays in a process that includes finding the weights of its features, determining appropriate scaling parameters, and assigning scores (Attali and Burstein 2006).

The weights of individual features can be determined by simply applying a multiple linear regression technique with the standardised human score as an outcome and standardised feature scores as predictors. However, the weights of individual features can also be determined by content experts or by setting them to values determined during prior similar assessments. Attali and Burstein (2006) found that judgement-based weights are not less efficient than optimal weights (found through regression analysis). With *e-rater*, it is also possible to combine optimal and judgement-based weights of features. Generally, once essays’ *e-rater* continuous scores are determined, they are transformed to a set of ordinal essay ratings. This study, however, used the continuous *e-rater* scores.

Procedure

The assessment of content in *e-rater* can be turned ‘off’ by eliminating the associated vectors (Features 7 and 8) in the regression equation for predicting essay scores. The procedure for this study was to have the *e-rater* determine the scores and the values of content-related and all other features in all essays for subsequent use in various analyses.

Analyses

As a first step in the analysis, an ANOVA was performed on *e-rater* predicted scores using genre and grade-levels as fixed factors to determine the potential impact of genre and grade level on essay scores. The ANOVA was followed by the Scheffe post-hoc comparisons across genre.

This was followed by a set of three regression analyses, one for each genre. In these analyses, the unstandardised *e-rater* essay score was used as the outcome, while the values of the *e-rater* features were used as predictors in a hierarchical fashion. First, all non-content features were entered as predictors. Then the content-related features were added to the predictor set. This hierarchical order of entering predictors into the model allowed an evaluation of the amount of variance accounted for by the content-related predictors by examining the change in the R^2value between the two models. In these analyses, the variable length (Feature 12) was excluded because it was highly and significantly correlated with five of the remaining eleven predictors. Additionally, the most current version of the *e-rater* does not use the essay length as one of its features (Attali and Burstein 2006).

For the regression models that included both content and non-content features, the ratio of the sum of the standardised regression coefficients for content-related features to the total sum of all standardised regression coefficients was found for each genre separately.

To aid in the interpretation of the results, three other regression models (one for each genre) were run. These models also used the *e-rater* essay score as the outcome, but only the content-related features as predictors.

Results

Table 5 shows the means and standard deviations for all of the variables used in the AES prediction equations. Essays averaged 231.17 (*SD* = 141.68) words with a relatively large standard deviation. This, in part, may reflect writing productivity by grade. Sixth-graders (*N* = 576) had a mean essay length of 188.69 (*SD* = 113.67), but 8th-graders (*N* = 1092) averaged 252.58 (*SD* = 149.69) words, $t(1666) = -.9.90$, $p < .001$.

The inter-correlation matrix for all *e-rater* predictor variables is displayed in Table 6. Essay length was significantly correlated with five of the remaining 11 elements, including style errors ($r = .56, p < .05$), required discourse elements ($r = .63, p < .05$), type/token ratio ($r = .50, p < .05$), vocabulary rated with a '6' ($r = .49, p < .05$), and least common words ($r = .54$, $p < .05$). *E-rater* had exact agreements with human raters of $r = .86$, $p < .01$ when the content variables were included, and $r = .85, p < .01$ when content was excluded from the prediction equation. The correlation between scores where content was included and the rescaled scores was $r = .99, p < .01$).

An ANOVA was performed on *e-raters'* predicted scores with content turned 'on' using grade and genre of writing as the independent variables. The results of this analysis are shown in Table 7.

This analysis is based on 648 essays written in response to persuasive prompts, 573 to expository, and 438 to descriptive prompts. It may be seen that there are significant main effects due to grade, $F(1,1653) = 58.71$, $p < .001$) and genre $F(2, 1653) = 20.57$, $p < .001$, but no significant interaction. The essay score means resulting from the ANOVA analysis are displayed graphically in Figure 1.

Scheffe post-hoc analyses were performed to look at pairwise contrasts across genre of writing. A summary is provided in Table 8. It can be observed that two of the three pairwise post-hoc comparisons were statistically significant, indicating that essays written in response to descriptive prompts, as a group, had a significantly higher mean score than the essays written in response to either expository or persuasive prompts.

Although the previous analysis shows that there are differences by grade and genre, it does not evaluate the impact of content by genre of writing. To address this question, the unstandardised multiple linear regression content model was compared with the unstandardised non-content model by genre

Table 9 shows the proportion of variance (R^2) accounted for by unstandardised regression models without content-related predictors (Model 1) and with content-related features (Model 2).

Table 5. The means and standard deviations for the AES predictor variables.

Variable	N	Min	Max	Mean	SD
Grammar	1668	.67	1.00	.99	.01
Usage	1668	.93	1.00	1.00	.01
Mechanics	1668	−.33	1.00	1.00	.06
Style	1668	.27	1.00	.85	.11
Required discourse elements	1668	−8.00	.00	−3.23	2.79
Length discourse elements	1668	3.00	385.00	43.48	30.27
Types/tokens	1668	.00	.71	.41	.12
Vocabulary with '6'	1668	.00	.23	.08	.03
Vocabulary	1668	1.00	6.00	3.94	1.43
Least common words	1659	20.70	90.60	58.57	8.30
Avg. word length	1668	2.33	6.17	3.97	.36
Length	1668	3.00	1168.00	231.17	141.68

Table 6. Correlations of the AES predictor variables.

	N = 1668, except those correlated with least common words where N = 1659											
Variable	G	U	M	S	RDE	LDE	T/T	V6	V	LCW	AWL	L
Grammar error (G)	1											
Usage error (U)	.07	1										
Mechanics errors (M)	.53*	.08	1									
Style errors (S)	.14	.05		1								
Required discourse elements (RDE)	.13	.07	.23	.45*	1							
Length discourse elements (LDE)	.05	−.04	.08	.17	−.32*	1						
Types/tokens (T/T)	.15	.03	.26	.02	.50*	.14	1					
Vocabulary with ‘6’ (V6)	.14	−.04	.27	.34*	.26	.25	.19	1				
Vocabulary (V)	.05	.01	.13	.17	.12	−.04	.02	.11	1			
Least common words (LCW)	.03	.08	.12	.39*	.40*	.17	.33*	.17	.05	1		
Avg. word length (AWL)	.09	.06	.13	.35*	.33*	−.09	.03	−.11	.28	.19	1	
Length (L)	.15	.07	.25	.56*	.63*	.29	.50*	.49*	.07	.54*	.23	1

* p < .05 (two-tailed); Bonferroni correction [(.05)/66 = .00075] made to preserve the Type I error rate of .05.

Table 7. Summary of the ANOVA on the *e-rater* predicted scores by grade (6 and 8) and genre (persuasive, expository, and descriptive.

Source	df	Mean square	F	Sig.
Corrected model	5	26.27	16.73	.000
Intercept	1	15340.68	9765.73	.000
Grade (Gr)	1	92.22	58.71	.000
Genre (Ge)	2	32.31	20.57	.000
Gr × Ge	2	1.00	.62	.536
Error	1653	1.57		
Total	1659			

$R^2 = .048$ (Adjusted $R^2 = .045$).

Table 8. Post-hoc comparisons of *e-rater* predicted score by genre.

					95% Confidence Interval	
(I) genre	(J) genre	Mean Difference (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
Expository	Persuasive	−.12	.07	.28	−.29	.06
	Descriptive	−.38*	.08	.00	−.57	−.18
Persuasive	Expository	−.12	.07	.28	−.06	.29
	Descriptive	−.27*	.08	.00	−.46	−.08
Descriptive	Expository	.38*	.08	.00	.18	.57
	Persuasive	.27*	.08	.00	.08	.46

Based on observed means.
* The mean difference is significant at the .05 level.

Table 9. R^2 model comparisons across genre for unstandardised regressions with content and without content.

					Change statistics				
Genre Model	R	R^2	Adj. R^2	Std. error of the estimate	R^2 change	F change	df1	df2	Sig. F change
Persuasive									
1	0.95	0.90	0.90	0.40	0.90	640.34	9	638	0.00
2	0.95	0.90	0.90	0.40	0.00	6.12	2	636	0.00
Expository									
1	0.96	0.92	0.92	0.38	0.92	681.25	9	563	0.00
2	0.96	0.93	0.92	0.36	0.01	33.00	2	561	0.00
Descriptive									
1	0.94	0.88	0.88	0.44	0.88	358.95	9	428	0.00
2	0.97	0.94	0.94	0.32	0.06	202.54	2	426	0.00

Figure 2 graphically displays the proportion of variance associated with content for the three writing-genres.

As a basis of comparison, the R^2 values for the three genre multiple linear regression equations using only the two predictor variables associated with content were .35 for persuasive, .38 for expository, and .58 for the descriptive genre.

Discussion

This study has determined the proportion of variance accounted for by content across different genres of writing. In earlier conceptualisations of essay evaluation and scoring, it had been

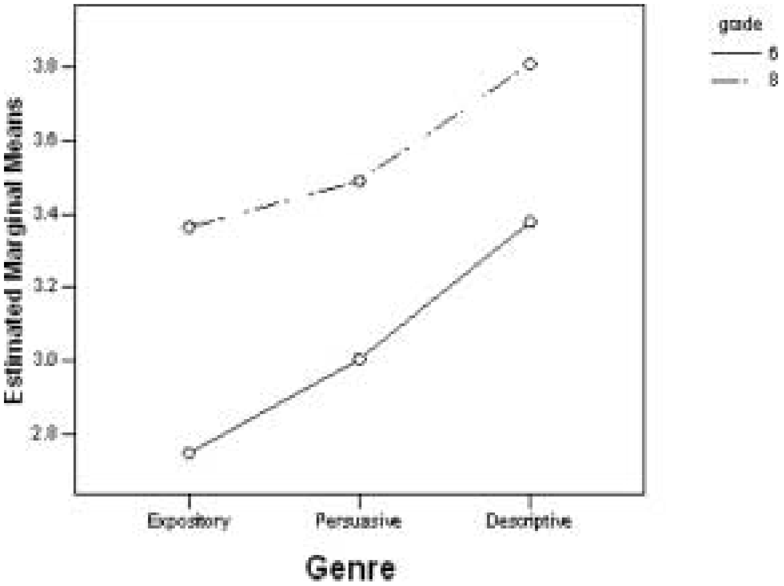

Figure 1. A plot of mean ratings of essays by grades 6 and 8 and genre.

assumed that content played a dominant or important part in assigning scores to prompt-specific essays (Landauer et al. 1998). In the current study, up to 58% of the variance could have been accounted for by knowledge of the two content-related variables. However, in the presence of the other predictor variables, *e-rater* scoring showed that content had relatively little weight in the prediction equations. For persuasive writing, it accounted for less than a half of 1%. For expository writing, it accounted for almost 1%, and for descriptive writing, it accounted for almost 6% of the additional variance in the *e-rater* score (after the non-content related predictors entered the model.) This order corresponded to the order of standardised regression weights assigned to the variables in the standardised egression equations for predicting *e-rater* scores for the three writing genres. The ratio of the summed standardised regression coefficients of the two content variables to the sum of the standardised weights for all predictors was sequenced as follows: Persuasive (.05), Expository (.09), and Descriptive (.23).

E-rater content-related features were composed of the variables Vocabulary and Vocabulary of Essays Scored with a '6'. The Vocabulary variable had relatively low correlations with the other predictors and Vocabulary of Essays Scored with a '6' had a significant correlation only with 'Style'. So, while the argument might be made that other predictors are masking the contributions embodied in the two content predictors, their lack of correlation with the other predictor variables would argue against this reasoning.

Evaluations of responses to prompt variations ran parallel with grade level. While one would expect 8th grade writing to be better than 6th grade writing, the results are a bit surprising because the different prompts used samples and raters unique to the prompt. The expectation for raters is that an average essay would be rated with a '3' in both groups. This expectation was supported for the 6th grade group, but the 8th grade ratings were significantly higher. In terms of genre, the mean ratings for both grade levels ran from lowest to highest for expository, persuasive, and descriptive genre. Descriptive essays had significantly higher mean ratings than persuasive or expository. The pattern order predicted (from lowest proportion of variance accounted for by

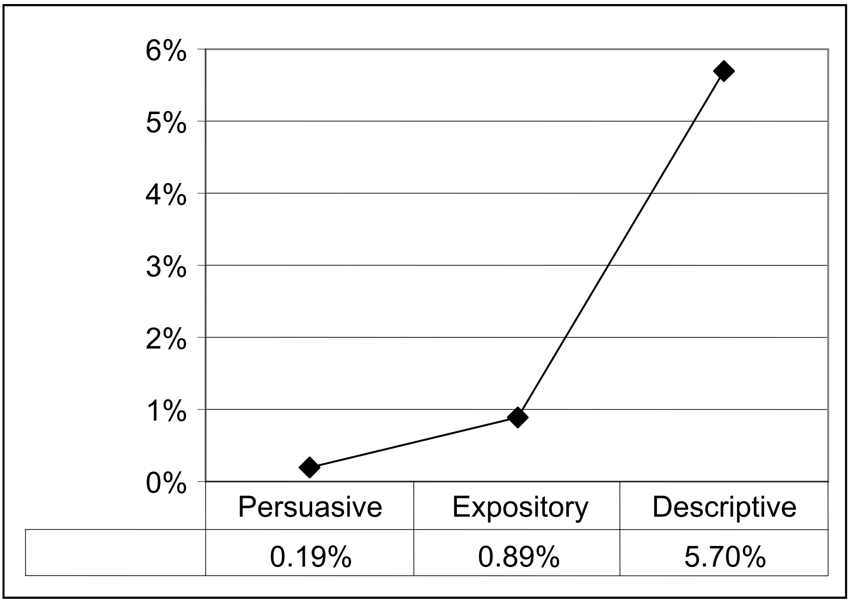

Figure 2. The proportion of variance associated with content for the three genres of writing.

content to highest proportion of variance accounted for by content) occurred in the predicted direction: persuasive, expository, and descriptive.

Because of the limited number of prompts across and within grade levels, the extent to which the sampling of prompts played a role in restricting the generalisations could not be determined precisely. For example, we ran the multiple regressions by grade. The pattern noted above fit particularly well for the 8th grade essays, but was not as strong for the 6th grade essays.

However, the results from this study certainly suggest some tentative conclusions: first, the proportion of variance accounted for by 'content', in the presence of other predictor variables, is not as high as had been previously stipulated; second, although prompt-level calibration for each grade makes intuitive sense, some upward 'creep' in the means for each group was apparent in spite of the fact that the scoring rubrics for all of the prompts were similar; third, there is some subjective bias in raters preference for essays written in response to some types of prompts over others.

Clearly when the purpose of the essay is to demonstrate mastery of some content domain, the notion that variance attributable to content might overlap with other predictors is a bit of a conundrum. Page (2003), in implementing Project Essay Grade, had the variables associated with content available in his essay scoring engine, but most of his studies were run with content turned 'off'. The use of content in this sense might be analogous to the social science researchers' view of sampling error. Much is made of it (probably because we have a relatively good understanding of how it works), but in most studies, sampling error accounts for less than 4% of variability in results. Most of the random variance in a typical study is more likely to be accounted for by measurement error than by sampling error. In automated essay scoring, we may attribute variability of scores due to content under the belief that content is better understood than some of the other criteria upon which essays are graded, but in fact, more emphasis may be given to these other attributes of writing.

In attempts to construct more generalisable essay models, the present study suggests that the effects related to at least two dimensions are worth further investigation: grade level and prompt genre. In fact this approach forms the basis of a proposed calibration approach outlined by Attali (2006). If general models can be configured to grade level and prompt type (this is referred to in their proposal as 'programme level'), then the sample sizes required to calibrate for a specific prompt may be significantly reduced. This, in turn, may lead to the greater availability of prompts for AES scoring. Ultimately it may lead to the ability of classroom teachers to create their own AES tailored prompts.

Finally we return to the notion that raters may have a preference for certain kinds of prompts. In the Shermis et al. (2001) study, practical and analytical prompts were rated significantly higher than traditional and emotional prompts. Although the current study's classification scheme did not parallel the genre generated in the original study, the conclusions may be the same: it is worth calibrating prompts to assign them a difficulty value, much in the same way that the execution of moves in a sporting event (e.g., diving) are weighted. An optimal score can be obtained by performing adequately in response to a difficult prompt compared to performing well on an easy prompt. AES holds considerable promise for understanding the calibration of prompts and its use in understanding the complexities of essay scoring by raters or by computers.

Acknowledgements

We would like to thank the Educational Testing Service for providing the data used in this study. The opinions expressed in this paper are those of the authors and may not necessarily represent opinions held by the Educational Testing Service. We are also indebted to Francis DiVesta and Don Powers for reviewing the manuscript and providing invaluable helpful comments.

Authors' note

This article is based on a paper presented at the annual meeting of the National Council of Measurement in Education, Montreal, Canada, April 12–14, 2005.

Notes

1. After Alexander Bain, although the mode of 'poetry' is not used here. We use the terms 'argument' and 'persuasive' interchangeably.
2. As of this writing, the current version of *e-rater* is 3.0.

Notes on contributors

Mark D. Shermis is a professor and chair in the Department of Educational Psychology at the University of Florida. His research interests are in automated essay scoring, the use of technology in assessment, and large-scale testing issues.

Alexandr Shneyderman is a director in the Department of Assessment, Research, and Data Analysis of the Miami-Dade County Public Schools system. His research interests include automated essay scoring, measurement, and evaluation studies in education.

Yigal Attali is a senior research scientist at the Educational Testing Service. His research interests are in automated scoring applications and the use of technology in assessment.

References

Attali, Y. 2004. Exploring the feedback and revision features of *Criterion*. Paper presented at the National Council on Measurement in Education, April 13–15, in San Diego, CA.

———. 2006. On-the-fly automated essay scoring. Paper presented at the National Council on Measurement in Education, April 8–10, in San Francisco, USA.

Attali, Y., and J. Burstein. 2006. Automated Essay Scoring With *e-rater* V.2. *Journal of Technology, Learning, and Assessment* 4, 3. http://www.jtla.org.

Burstein, J. 2003. The *e-rater* scoring engine: Automated essay scoring with natural language processing. In *Automated essay scoring: A cross-disciplinary perspective,* ed. M.D. Shermis and J. Burstein, 113–122. Mahwah, NJ: Lawrence Erlbaum Associates.

Burstein, J., M. Chodorow, and C. Leacock. 2003. Criterion: Online essay evaluation: An application for automated evaluation of test-taker essays. Paper presented at the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence, August 12–14, in Acapulco, Mexico.

D'Angelo, F.J. 1976. Modes of discourse. In *Teaching composition: 10 bibliographical essays,* ed. G. Tate, 111–135. Fort Worth, TX: Christian University Press.

Elliot, S. 2003. Intellimetric: From here to validity. In *Automated essay scoring: A cross-disciplinary perspective,* ed. M.D. Shermis and J. Burstein, 71–86. Mahwah, NJ: Lawrence Erlbaum Associates.

Engelhard, G. 1992. The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education* 5, no. 3: 171–191.

Ericsson, P.F., and R. Haswell, eds. 2006. *Machine scoring of student essays: Truth and consequences.* Logan, UT: Utah State University Press.

Hamp-Lyons, L., and S.P. Mathias. 1994. Examining expert judgements of task difficulty on essay tests. *Journal of Second Language Writing* 3, no. 1: 49–68.

Kawate-Mierzejewska, M. 2003. *E-rater* software. Paper presented at the Japanese Association for Language Teaching monthly chapter meeting, March 23, in Tokyo, Japan.

Keith, T.Z. 2003. Validity and automated essay scoring systems. In *Automated essay scoring: A cross-disciplinary perspective,* ed. M.D. Shermis and J. Burstein, 147–168. Mahwah, NJ: Lawrence Erlbaum Associates.

Landauer, T.K., P.W. Foltz, and D. Laham. 1998. Introduction to latent semantic analysis. *Discourse Processes* 25, nos. 2–3: 259–284.

Landauer, T.K., D. Laham, and P.W. Foltz, 2003. Automated scoring and annotation of essays with the Intelligent Essay Assessor. In *Automated essay scoring: A cross-disciplinary perspective,* ed. M.D. Shermis and J. Burstein, 87–112. Mahwah, NJ: Lawrence Erlbaum Associates.

Leacock, C., and M. Chodorow. 2003. C-rater: Scoring of short-answer questions. *Computers and the Humanities* 37, no. 4: 389–405.

Leki, I., and J. Carson. 1997. Completely different worlds: EAP and the writing experiences of ESL students in university courses. *TESOL Quarterly* 31, no. 1: 39–70.

McTavish, D.G., K.C. Litkowski, and S. Schrader 1997. A computer content analysis approach to measuring social distance in residential organisations for older people. *Social Science Computer Review* 15, no. 2: 170–180.

Page, E.B. 2003. Project Essay Grade: PEG. In *Automated essay scoring: A cross-disciplinary perspective,* ed. M.D. Shermis and J. Burstein, 43–54. Mahwah, NJ: Lawrence Erlbaum Associates.

Page, E.B., T. Keith, and M.J. Lavoie. 1995. Construct validity in the computer grading of essays. Paper presented at the annual meeting of the American Psychological Association, August 11–15, in New York, USA.

Page, E.B., and N.S. Petersen. 1995. The computer moves into essay grading: Updating the ancient test. *Phi Delta Kappan* 76, no. 7: 561–565.

Rajecki, D.W., J.A. Dame, K.J. Creek, P.J. Barrickman, C.A. Reid, and D.C. Appleby. 1993. Gender casting in television toy advertisements: Distributions, message content analysis, and evaluations. *Journal of Consumer Psychology* 2, no. 2: 307–327.

Ruth, J., and S.M. Murphy. 1988. *Designing writing tasks for the assessment of writing.* Norwood, NJ: Ablex.

Shermis, M.D., J. Burstein, and C. Leacock. 2006. Applications of computers in assessment and analysis of writing. In *Handbook of writing research,* ed. C.A. MacArthur, S. Graham and J. Fitzgerald, 403–416. New York: Guilford Publications.

Shermis, M.D., J.L. Rasmussen, D.W. Rajecki, J. Olson, and C. Marsiglio. 2001. All prompts are created equal, but some prompts are more equal than others. *Journal of Applied Measurement* 2, no. 2: 154–170.

Shermis, M.D., A. Shneyderman, and Y. Attali. 2005. *How important is content in the ratings of essay assessments?* Paper presented at the Paper presented at the annual meeting of the National Council of Measurement in Education, April 12–14, in Montreal, Canada.

Vantage Learning. 2001. *A preliminary study of the efficacy of IntelliMetric™ for use in scoring Hebrew assessments.* Newtown, PA: Vantage Learning.

———. 2002. *A study of IntelliMetric™ scoring for responses written in Bahasa Malay* (No. RB-735). Newtown, PA: Vantage Learning.

Weigle, S.C. 1998. Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing* 6, 2: 145–178.

Wolfe, E.W. 2004. Identifying rater effects using latent trait models. *Psychology Science* 46, no. 1: 35–51.