

Evidence for the Interpretation and Use of Scores from an Automated Essay Scorer

Paul Nichols

PEM Research Report 05-02

March 2005



*Using testing and
assessment to
promote learning*

Pearson Educational Measurement has been a trusted partner in district, state and national assessments for more than 50 years. As a pioneer and the largest comprehensive provider of educational assessment products, services and solutions, PEM helps states and large school districts meet the requirements of education reform while using testing and assessment to promote learning.

PEM Research Reports provide dissemination of PEM research and assessment-related articles prior to publication. PEM reports in .pdf format may be obtained at:

<http://www.pearsonedmeasurement.com/research/research.htm>



Abstract

This paper examined validity evidence for the scores based on the Intelligent Essay Assessor (IEA), an automated essay-scoring engine developed by Pearson Knowledge Technologies. A study was carried out using the validity framework described by Yang, et al. (2002). This framework delineates three approaches to validation studies: examine the relationship among scores given to the same essays by different scorers, examine the relationship between essay scores and external measures, and examine the scoring processes used by the IEA.

The results of this study indicated that, although relationships among scores given to the same essays by different scorers (percent agreement, Spearman rank-order correlation, kappa statistic and Pearson correlation) indicated a stronger relationship between two human readers than between the IEA and a human reader, stronger relationships were found between the IEA and experts than between readers and experts. In addition, the results of examining the scoring processes used by the IEA showed that the IEA used processes similar to a human scorer. Furthermore, the IEA scoring processes were more similar to processes used by proficient human scorers than to processes used by non-proficient or intermediate human scorers.

The results of this study provided positive evidence for the use of IEA scores as measures of writing achievement. Further research with the IEA in other assessments and grade levels will be helpful in generalizing the results of this study and further strengthening the validity of IEA for scoring writing assessments.

Introduction

Essay assessment can be expensive and time-consuming (Hardy, 1996; Khattri & Sweet, 1996; Parkes, 2000). An essay test is a more expensive and time consuming alternative compared to multiple-choice tests if the essay test is held to the same level of domain coverage and reliability as the multiple-choice test (Hoover & Bray, 1995; Reckase, 1995). Most of the cost and time associated with essay assessment is attributed to scoring (White, 1986).

The result of the relatively greater expense and time associated with scoring essays is that, compared to multiple-choice tests, fewer tasks tend to be administered and feedback on test takers' performance is delayed. The administration of relatively fewer tasks compared to the number of items on a multiple-choice test leads to at least two problems. First, the domain assessed by an essay test is more narrow than the domain assessed by a multiple-choice test, all other things being equal across essay and multiple choice tests (Hoover & Bray, 1995). Second, the generalizability of the test scores from an essay test is more restricted compared to test scores from a multiple-choice test, all other things being equal.

The delay in the return of feedback on essay test performance compared to multiple-choice tests lessens the usefulness of the results for teachers and administrators. The importance of timely feedback has been documented in detailed studies of the relationship between large-scale assessment and classroom and school practices (Thacker, Koger & Koger, 1998). The lack of timely feedback diminishes the influence of test results on educational practice (Haertel, 1999).

Computer-automated scoring for essay responses, or automated essay-scoring (AES), may potentially reduce the expense and time associated with scoring essay responses. Automated essay scoring programs are based on varying combinations of artificial intelligence, computational linguistics and cognitive science [see Clauser, Kane and Swanson (2002) and Yang, Buckendahl, Juskiewicz and Bhola (2002) for a review of AES]. An AES can score a response and return feedback in seconds. Furthermore, an AES can score twenty-four hours-a-day, seven days a week. The greatest limitation of AES is that essay responses must be key-entered before they are processed by an AES.

As with the use of human judges in scoring essay responses, validity is a primary concern in the use of AES to score essay responses. As Clauser, Kane and Swanson (2002) point out, the

criteria for evaluating the use of AES are the same as the criteria for evaluating the use of human judges in scoring essay responses. However, “different aspects of the validity argument may require emphasis as a function of the scoring procedure” (Clauser, Kane, & Swanson, 2002; p. 414).

Much of the past validity evidence offered in the use of AES to score essay responses has come in the form of correspondence between scores produced by the use of AES and scores produced by human judges. Correspondence may be measured as the percentage of agreement between scores produced by AES and scores produced by human judges. Correspondence may also be measured using an index that corrects for chance agreement, such as Cohen’s Kappa. In addition, correspondence may be explored by investigating sources of variance via generalizability analysis.

Furthermore, much of the validity evidence offered in support of AES has come from developers of AES programs. For example, the developers of the Intelligent Essay Assessor (IEA) offered by Pearson Knowledge Technologies have published much of the evidence supporting the validity of the IEA (i.e., Foltz, Laham & Landauer, 1999). Similarly, the developers of e-Rater offered by Educational Testing Service have published much of the evidence supporting the validity of E-rater (i.e., Burstein, Kukich, Wolff, Lu & Chodrow, 1998; Burstein, Marcu, Andreyev, & Chodorow, 2001). Such a phenomenon is expected early in the introduction of a new technology.

This paper presents validity evidence for the meaning and uses of scores produced by one AES: the IEA offered by Pearson Knowledge Technologies. Building on past AES validity research, this paper organizes the evidence using the framework described by Yang, Buckendahl, Juskiewicz, and Bhola (2002). The following three general approaches to AES validation studies are described by Yang, et.al., (2002): presenting the relationship among scores given to the same essays by different scorers, presenting the relationship between essay scores and external measures, and examining the scoring processes used by the AES. In this paper, evidence from each of these approaches is used in constructing an interpretive argument around the use of IEA scores.

Method

In this study, the constructed responses of examinees from three grade levels to writing prompts were scored by scoring directors who served as experts, trained readers and by the IEA.

Participants

A total of 3244 test takers from a large Southern state participated in this study: 1102 test takers from grade 4, 1311 test takers from grade 8, and 831 test takers from grade 10. Each student responded to only one essay writing prompt as part of a large and ongoing state assessment program.

Prompts

The prompts used in this study were two prompts from grade 4, two prompts from grade 8, and a prompt from grade 10. The grade 4 prompts were an expository and a narrative writing prompt. The grade 8 prompts were an expository and a persuasive writing prompt. The grade 10 prompt was an expository writing prompt. The numbers of essay responses for each prompt are shown in Table 1.

The original essay responses were hand-written. For the current study, the hand-written responses were key-entered and provided to vendors in electronic format. A 100% recheck was done on all key entry.

Table 1. *Number of essay responses scored for each prompt*

Prompt	N
Grade 4 Narrative	553
Grade 4 Expository	549
Grade 8 Persuasive	693
Grade 8 Expository	618
Grade 10 Expository	831
Total	3244

Readers

This study included both trained readers and experts. A total of 53 readers scored essay responses. Readers were trained together using a common curriculum. Experts were two Pearson Educational Measurement scoring directors from the Performance Scoring Center.

Procedures

The essay responses were scored three times: by experts, by readers, and by the IEA. The procedures used by each scoring group are described below.

Experts. Initially, each response was scored independently by two experts using paper copies of the response. Responses were scored using a six-point rubric. For this study, all responses for which the experts' scores did not agree were removed from the data set. Note that the experts were in exact agreement for 96% of the responses. In addition, all responses for which a code of A, indicating a blank response, or C, indicating an illegible response, were assigned by experts were removed from the data set. A blank response would provide no information on the performance of the scoring software and an illegible response could not be key-entered.

Therefore, scores from experts ranged from 1 through 6 to B, representing off-task responses, or D, representing foreign language responses. The resulting set of essays to be scored by vendors consisted of 3,244 essay responses.

Readers. The set of 3,244 essay responses were then randomly assigned to readers to score. Each essay was scored by two readers using paper copies of the response. Readers scored responses using the same six-point rubric used by experts.

IEA. The IEA was "trained" to score responses to each prompt. To train these programs, a set of training and a set of cross-validation essay responses are used for each prompt. Using the training set, the essay-scoring programs "learn" how to score the essay responses by matching characteristics of the essay responses with the scores these essay responses received. The scoring criteria the essay-scoring programs have "learned" are then checked against scores given to the set of cross-validation essay responses. Generally, the larger the number of essay responses used in training and cross-validation sets the more accurate the automated essay-scoring programs.

For each prompt, the IEA was trained using a relatively small set of 222 essay responses that were divided between training and cross-validation sets. Pearson Knowledge Technologies decided how to divide these 222 sets between training and validation. All essay responses used to train the IEA were first scored by readers from the PEM Performance Scoring Center. These essay responses were used to train the IEA but were excluded from the study itself.

A goal of this study was to evaluate the impact of the size of the training set and the cross-validation set of essay responses used to “train” the automated essay scoring programs. To accomplish this goal, the grade 10 expository writing prompt was scored twice. Initially, the automated essay-scoring software was trained using a relatively small set of 222 essay responses that were divided between training and cross-validation sets. Pearson Knowledge Technologies decided how to divide these 222 sets between training and validation.

After scoring the essay responses, the IEA was trained again using a larger training set and cross-validation set of essay responses. A total of 417 sets of essay responses and scores were in the larger training set and cross-validation set of essay responses. The IEA used 227 essay responses in the training set and 190 papers in the cross-validation set. The essay responses were then scored a second time.

A total of 831 essay responses were scored after the software had been trained using a relatively small training set and cross-validation set of essay responses. A total of 636 essay responses were scored after the software had been trained using a relatively large training set and cross-validation set of essay responses. More essays were scored after training using relatively small training and cross-validation sets because more essays were available to score.

Results

The presentation of the results of this study is organized around the three general approaches to validation evidence for AES described by Yang, et.al., (2002): presenting the relationship among scores given to the same essays by different scorers, presenting the relationship between essay scores and external measures, and examining the scoring processes used by the IEA.

Relationship Among Scores Given to the Same Essays by Different Scorers

Evidence of relationships among scores given to the same essays by different scorers may be provided by analyses of the level of agreement between scores assigned by readers and scores produced by the IEA for the same papers. Indexes of this relationship are percent agreement, Cohen's Kappa, and Spearman's rank order coefficient.

Descriptive statistics for scores assigned by readers and scores produced by the IEA are shown in Table 2. Table 2 provides summary descriptive statistics overall and by prompt for IEA scores and readers. Note that data from all essay responses flagged by IEA, or given a B or D rating by readers were removed before descriptive statistics were computed. As Table 2 shows, there was little difference in mean scores between the IEA, reader 1 and reader 2. None of the differences between means were statistically significant.

Table 2. *Descriptive Statistics for Reader and IEA Scores*

Source	N	Mean	SD	Median	Min	Max
Grade 4 Narrative						
IEA	500	3.55	0.80	4	1	5
Reader 1	500	3.33	0.92	3	1	6
Reader 2	500	3.31	0.97	3	1	6
Grade 4 Expository						
IEA	514	3.09	0.81	3	1	5
Reader 1	514	3.19	1.01	3	1	5
Reader 2	514	3.12	0.99	3	1	5
Grade 8 Persuasive						
IEA	650	3.31	0.83	3	1	5
Reader 1	650	3.24	1.05	3	1	6
Reader 2	650	3.14	0.96	3	1	6
Grade 8 Expository						
IEA	584	3.60	0.83	4	1	6
Reader 1	584	3.86	0.96	4	1	6
Reader 2	584	3.92	1.01	4	1	6
Grade 10 Expository						
IEA	618	3.72	1.10	4	1	6
Reader 1	618	3.60	1.12	4	1	6
Reader 2	618	3.68	1.05	4	1	6
Across Prompts						
IEA	2866	3.46	0.91	3	1	6
Reader 1	2866	3.45	1.05	3	1	6
Reader 2	2866	3.44	1.05	3	1	6

Percent agreement. Percent agreements between scores from the IEA and readers' scores ordinary scorers are shown in Table 3. The essay responses flagged as unscorable by IEA, or given a B or D rating by readers, were set to 0 for this analysis.

Table 3. *Percent agreement by prompt and across prompts between IEA and reader 1, IEA and reader 2, and reader 1 and reader 2 (N=3244).*

Prompt	N	IEA/Reader 1		IEA/Reader 2		Reader 1/2	
		Exact	Exact + Adjacent	Exact	Exact + Adjacent	Exact	Exact + Adjacent
Grade 4 Narrative	553	47.20	93.67	44.67	94.58	52.80	94.03
Grade 4 Expository	549	49.91	95.26	51.37	95.45	53.01	93.99
Grade 8 Persuasive	693	50.36	94.81	48.48	95.09	47.19	92.06
Grade 8 Expository	618	45.63	91.26	42.88	89.48	41.26	91.26
Grade 10 Expository	831	41.16	91.21	45.85	91.70	43.32	89.29
Across prompts	3244	46.48	93.09	46.58	93.13	47.01	91.86

As the data in Table 3 indicate, exact agreement tended to be higher between reader 1 and reader 2 than between the IEA and reader 1 or the IEA and reader 2. When adjacent scores are included, rate of agreement tends to be higher between the IEA and reader 1 or the IEA and reader 2 than between reader 1 and reader 2.

Spearman rank-order correlation. In addition to percent agreement, the Spearman rank-order correlation between IEA scores and readers' scores was used as a measure of correspondence. The Spearman rank-order correlation is a measure of the strength of the linear association based on the rank of the data values. To compute the Spearman rank-order correlation, both IEA scores and readers' scores were transformed into ranks so that essays assigned a 1 were treated as the lowest ranking score and essays assigned a 6 were treated as the highest ranking category. In computing the Spearman correlation, all essay responses with reader ratings of "B" or "D" or flagged as unscorable by IEA were excluded from this analysis. The Spearman rank-order correlations between IEA scores and readers' scores are shown in Table 4.

As Table 4 shows, a clear pattern emerged in the results. For each prompt, the Spearman rank-order correlation was higher between reader 1 and reader 2 than between the IEA and either

reader 1 or reader 2. Across all prompts, the finding was the same. However, these differences in the Spearman rank-order correlation were small.

Table 4. *Spearman correlation by prompt and across prompts between IEA and reader 1, IEA and reader 2, and reader 1 and reader 2*

	N	IEA/Reader 1	IEA/Reader 2	Reader 1/2
Grade 4 Narrative	548	0.5589	0.5859	0.6207
Grade 4 Expository	545	0.6194	0.6121	0.6472
Grade 8 Persuasive	692	0.6271	0.6001	0.6208
Grade 8 Expository	613	0.5160	0.5098	0.5242
Grade 10 Expository	827	0.6388	0.6372	0.6412
Total	3225	0.6038	0.6026	0.6379

Kappa statistic. The fourth measure of correspondence comparing the IEA scores to readers' scores was the kappa statistic. The kappa statistic is a measure of the agreement between two readers. If the ratings of the two readers are viewed as two independent ratings of the papers, the kappa coefficient equals +1 when there is complete agreement between the two readers. When the agreement between the two readers exceeds chance agreement, kappa is positive, with its magnitude reflecting the strength of agreement. Although this is unusual in practice, kappa is negative when the agreement between the two readers is less than chance agreement.

The kappa statistic adjusts the measure of reliability for chance agreement. In contrast, percent agreement may provide an inflated measure of reliability because readers will agree on some papers by chance even if randomly assigning scores. Note that readers' scores of B and D were replaced with 0 and the IEA scores for papers assigned flags were replaced with 0 in the analysis using the kappa statistic. As shown in Table 6, kappa values could be computed for only three prompts. Kappa values could not be computed when the matrix of IEA scores and readers' scores was not symmetric (i.e., the matrix had same number of rows as columns).

All the kappa values between the IEA scores and readers' scores were well above the chance level value of 0.00. Note that no clear pattern was present for the results. For each prompt, sometimes the Kappa value was higher between reader 1 and reader 2 than between the

IEA and either reader 1 or reader 2 but sometimes the reverse was found. Across all prompts, the Kappa value was higher between reader 1 and reader 2 than between the IEA and either reader 1 or reader 2.

Table 5. *Kappa statistics by prompt and across prompts between IEA and reader 1, IEA and reader 2, and reader 1 and reader 2*

	N	IEA/Reader 1	IEA/Reader 2	Reader 1/2
Grade 4 Narrative	548	NA	NA	0.3287
Grade 4 Expository	549	0.2774	0.2965	0.3495
Grade 8 Persuasive	692	NA	NA	NA
Grade 8 Expository	618	0.2240	0.1926	0.1826
Grade 10 Expository	831	0.2167	0.2681	0.2534
Total	3244	0.2562	0.2557	0.2880

Pearson Correlation. The Pearson correlation between the IEA scores and readers' scores was used as another measure of correspondence. The Pearson correlation measures the direction and degree of the relationship between pairs of scores. The Pearson correlation is appropriate only when both variables lie on an interval scale. Consequently, all essay responses with reader ratings of "B" or "D" or scores flagged by the IEA were excluded from this analysis.

The pattern of results shown in Table 6 is not as strong as the pattern of results for the Spearman rank-order correlation. The Pearson correlation for each prompt was sometimes higher between reader 1 and reader 2 than between the IEA and either reader 1 or reader 2. Across all prompts, the Pearson correlation was higher between reader 1 and reader 2 than between the IEA and either reader 1 or reader 2. As with the other indicators of agreement, these differences in the Pearson correlations were small.

Table 6. *Pearson correlation by prompt and across prompts between IEA and reader 1, IEA and reader 2, and reader 1 and reader 2*

	N	IEA/Reader 1	IEA/Reader 2	Reader 1/2
Grade 4 Narrative	548	0.5572	0.5799	0.6271
Grade 4 Expository	545	0.6513	0.6414	0.6798
Grade 8 Persuasive	692	0.6614	0.6271	0.6223
Grade 8 Expository	613	0.5358	0.5307	0.5525
Grade 10 Expository	827	0.6469	0.6396	0.6473
Total	3225	0.6269	0.6197	0.6509

Summary. These results show that scores assigned by the IEA tended to be similar in level to the scores assigned by readers. However, the scores assigned by the IEA were less variable than scores assigned by readers. In addition, the agreement between the IEA scores and either reader was lower than the agreement between the two readers. Finally, the IEA scores tended to be in a similar order as readers' scores as indicated by the moderately high Pearson and Spearman correlations.

Relationships Between Essay Scores and External Measures

Evidence of relationships between essay scores and external measures may be provided by measures of the agreement between scores produced by the IEA and scores assigned essays by experts. The expert scores are considered external measures because the IEA is trained using scores from readers rather than experts. Thus, expert scores are another, but independent, measure of achievement.

Descriptive statistics. Table 7 provides summary descriptive statistics overall and by prompt for IEA scores, readers' scores, and experts' scores. Note that data from all essay responses flagged by the IEA, or given a B or D rating by readers or experts were removed before descriptive statistics were computed. As Table 7 shows, mean and median IEA scores were generally similar to readers' scores but higher than experts' scores as shown by the difference in mean scores across prompts. No clear pattern emerged for individual prompts.

Table 7. *Descriptive statistics for expert and IEA scores*

Source	N	Mean	SD	Median	Min	Max
Grade 4 Narrative						
IEA	547	3.52	0.80	4	1	5
Reader 1	547	3.30	0.93	3	1	6
Reader 2	547	3.29	0.97	3	1	6
Expert	547	3.41	0.58	3	1	6
Grade 4 Expository						
IEA	544	3.07	0.82	3	1	5
Reader 1	544	3.17	1.02	3	1	5
Reader 2	544	3.09	1.00	3	1	5
Expert	544	3.11	0.78	3	1	5
Grade 8 Persuasive						
IEA	692	3.30	0.84	3	1	5
Reader 1	692	3.23	1.06	3	1	5
Reader 2	692	3.12	0.97	3	1	5
Expert	692	3.03	0.83	3	1	5
Grade 8 Expository						
IEA	612	3.58	0.85	4	1	6
Reader 1	612	3.85	0.97	4	1	6
Reader 2	612	3.90	1.02	4	1	6
Expert	612	3.54	0.77	3	1	6
Grade 10 Expository						
IEA	826	3.58	0.98	4	1	6
Reader 1	826	3.60	1.17	4	1	6
Reader 2	826	3.64	1.12	4	1	6
Expert	826	3.62	0.91	4	1	6
Across Prompts						
IEA	3221	3.42	0.89	3	1	6
Reader 1	3221	3.44	1.07	3	1	6
Reader 2	3221	3.42	1.07	3	1	6
Expert	3221	3.36	0.83	3	1	6

Note: Excludes data from all essay responses flagged as off topic by IEA or given a B or D rating by experts or readers.

In addition, IEA scores, readers' scores and experts' scores differed in variability. Scores from readers were the most variable and scores from experts were the least variable as shown by the difference in the standard deviations across prompts. For individual prompts, scores from experts were the least variable for all five prompts and scores from readers were the most variable for all five prompts.

In summary, scores from the IEA and readers tended to be higher than experts' scores. However, scores from readers were more variable than scores for the IEA or experts.

Percent agreement. Percent agreement between the scores from the IEA and experts' scores was used as a measure of correspondence. Percent agreement was examined overall across prompts and then separately for each prompt. Percent agreement for each prompt was examined to test if the overall pattern of percent agreement held at the prompt level.

The frequency of agreement between scores from the IEA and experts' scores is shown in Table 8. The essay responses flagged as unscorable by IEA, or given a B or D rating by readers or experts, were set to 0 for this analysis. Overall, the IEA had a higher rate of exact agreement than either the first or second reader. The same pattern was evident for individual prompts.

When adjacent agreement is included along with exact agreement, the rate was again higher for the IEA than for either the first or second reader. The same pattern of agreement was evident for individual prompts.

Table 8. *Percent Agreement by Prompt and Across Prompts between IEA and Expert, Reader 1 and Expert, and Reader 2 and Expert (N=3244)*

Prompt	N	IEA/Expert		Reader 1/Expert		Reader 2/Expert	
		Exact	Exact + Adjacent	Exact	Exact + Adjacent	Exact	Exact + Adjacent
Grade 4 Narrative	553	64.74	98.92	50.81	96.38	51.72	95.66
Grade 4 Expository	549	67.21	99.09	58.47	96.36	53.73	97.45
Grade 8 Persuasive	693	49.21	94.66	49.06	91.34	49.21	94.52
Grade 8 Expository	618	61.17	98.06	47.09	92.56	47.57	89.48
Grade 10 Expository	831	58.72	97.83	46.81	93.02	49.34	94.22
Across prompts	3244	59.62	97.60	50.00	93.71	50.12	94.17

Spearman rank-order correlation. In computing the Spearman rank-order correlation shown in Table 9, all essay responses with expert ratings of “B” or “D” or flagged as unscorable by the IEA were excluded from this analysis. Across prompts, the Spearman rank-order correlation between IEA scores and experts' scores was higher than the Spearman rank-order correlation between readers' scores and experts' scores. For each prompt, the Spearman rank-order correlation between IEA scores and experts' scores was again higher than the Spearman

rank-order correlation between readers' scores and experts' scores. The only exception was the Grade 8 persuasive prompt.

Table 9. *Spearman Correlation by Prompt and Across Prompts between IEA and Experts, Reader 1 and Experts, and Reader 2 and Experts*

	N	IEA/Expert	Reader 1/Expert	Reader 2/Expert
Grade 4 Narrative	547	0.6602	0.5349	0.5837
Grade 4 Expository	544	0.7063	0.6776	0.6370
Grade 8 Persuasive	692	0.5468	0.5879	0.6021
Grade 8 Expository	612	0.6526	0.5101	0.5165
Grade 10 Expository	826	0.7353	0.6603	0.6706
Total	3221	0.6679	0.6118	0.6291

Kappa statistic. Note that when computing the kappa statistic the experts' scores of B and D were replaced with 0 and essay responses flagged as unscorable by the IEA were assigned a score of 0. As Table 10 shows, the kappa value was well above the chance level value of 0.00. The IEA had a higher kappa value than either the first or second reader when ratings are collapsed across prompts. The same pattern was evident for individual prompts.

Note that kappa values could be computed for only three of the five prompts. Kappa values could not be computed if the matrix of scores was not symmetric (i.e., the matrix had same number of rows as columns).

Table 10. *Kappa Statistics by Prompt and across Prompts between IEA and Experts, Reader 1 and Experts, and Reader 2 and Experts*

	N	IEA/Expert	Reader 1/Expert	Reader 2/Expert
Grade 4 Narrative	553	NA	0.2357	0.2659
Grade 4 Expository	549	0.4783	0.3878	0.3173
Grade 8 Persuasive	693	0.2570	NA	NA
Grade 8 Expository	618	0.4054	0.2309	0.2541
Grade 10 Expository	831	0.4095	0.2819	0.2995
Total	3244	0.3948	0.2968	0.2968

Pearson correlation. Table 11 shows the Pearson correlations computed between IEA scores and experts' final scores and between readers' scores and experts' scores. All essay responses with expert ratings of "B" or "D" and essay responses flagged as unscorable by the IEA were assigned a score of 0.

Table 11. *Pearson Correlation by Prompt and across Prompts between IEA and Experts, Reader 1 and Experts, and Reader 2 and experts*

	N	IEA/Expert	Reader 1/Expert	Reader 2/Expert
Grade 4 Narrative	547	0.6592	0.5566	0.6008
Grade 4 Expository	544	0.7431	0.7083	0.6792
Grade 8 Persuasive	692	0.5721	0.5977	0.6045
Grade 8 Expository	612	0.6843	0.5540	0.5412
Grade 10 Expository	826	0.7417	0.6783	0.6859
Total	3221	0.6924	0.6405	0.6479

Across prompts, the Pearson correlation between IEA scores and experts' scores was higher than the Pearson correlation between readers' scores and experts' scores. As was the finding using Spearman rank-order correlations, the Pearson correlation between IEA scores and experts' scores was again higher than the Pearson correlation between readers' scores and experts' scores for individual prompts. The only exception was the Grade 8 persuasive prompt.

Summary. These results show that the scores assigned by the IEA were generally higher and more variable than experts' scores. Across prompts, the IEA mean score was higher than the experts' mean score and the standard deviation of IEA scores was larger than the standard deviation of experts' scores. However, mean IEA scores were similar to the mean scores for readers. Furthermore, the variability of the IEA scores was closer to the variability of experts' scores than was the variability of readers' scores.

The IEA scores tended to be in a similar order as experts' scores as indicated by the moderately high Pearson and Spearman correlations. The order of scores assigned by the IEA tended to follow more closely the experts' scores than did readers' scores. In addition, the IEA scores agreed more closely with experts' scores than did readers' scores. This result was found across prompts as well as for each individual prompt.

Scoring Processes Used by the AES

Insight into the scoring processes used by the AES may be provided by examining how the IEA assigns scores to text. Does the IEA assign scores in ways consistent with the manner in which a human assigns scores? This section offers a comparison between the scoring process used by the IEA and empirical research about how human readers assign scores to essays.

The IEA. The IEA uses a statistical combination of several measures to produce a score for an essay (Landauer, Laham & Foltz, 2001). To compute a score, the IEA combines the following three meta-variables: content, style and mechanics. In addition, the IEA uses validity and confidence measures. Each meta-variable is a composite of two or more sub-components. Each meta-variable and subcomponent has an associated weight but only the weights of meta-variables are adjusted across essays. By default, the three meta-variables are combined using multiple regression on scores assigned essays by human raters as part of a training sample. The multiple-regression equation is constrained so that the content meta-variable is always assigned the greatest weight.

The greatest difference between the IEA and other AES programs, such as e-rater offered by Educational Testing Service (Burstein, Kukich, Wolff, Lu & Chodrow, 1998), is the use by the IEA of Latent Semantic Analysis (LSA). LSA is a machine-learning model of the human understanding of text. The content meta-variable uses LSA to predict the grade human scorers would have given an essay. LSA is a method for determining the similarity of meaning of words and passages by analysis of a large sample of machine readable language (Landauer, Foltz & Laham, 1998). LSA provides approximate estimates of the contextual usage substitutability of words in larger text segments, and of the kinds of-as yet incompletely specified-meaning similarities among words and text segments that such relations may reflect. The meanings of words and passages are induced from the analysis of text alone. However, LSA's understanding of the meaning of words and passages has been characterized by Landauer, Foltz & Laham (1998) as "analogous to a well-read nun's knowledge of sex" (p. 5).

The procedure for estimating the value for the content meta-variable follows six steps (Landauer, Laham & Foltz, 2001).

1. LSA is applied to extensive background machine readable text, typically before any essays are processed, to represent the meaning of words as used in the assessment domain.
2. A representative sample of essays, the training sample, is scored by human judges.
3. All training and to-be-scored essays are represented as an LSA vector.
4. Similarity is computed between each of the to-be-scored essays and the training essays.

5. Select a small number of training essays with which to compare the to-be-scored essay.
6. Compute a prediction of the score that the to-be-scored essay would have been given by using this small number of selected training essays.

The style meta-variable consists of a number of sub-components including the following:

- An index of the conceptual coherence among parts such as words, sentences and paragraphs measured as the cosine between vectors for parts to each other or the whole;
- An index of word flow measured as the degree to which previous words predict following words;
- An index of grammaticality estimated as the degree of normality of a sentence's grammatical structure relative to a corpus of good writing on the topic;
- Other indexes such as word usage and standard readability measures;

The mechanics meta-variable is a combination including spelling and punctuation.

In addition, the IEA reports a confidence flag based on the number of training essays that were similar to the to-be-scored essay.

A human reader. Several theories of readers' cognition during essay scoring have been described over the last several decades including the theories of Freedman and Calfee (1983), Frederiksen (1992), and Wolfe (1997). In this section, the theory proposed by Wolfe is reviewed and subsequently compared to the scoring processes used by the IEA.

The theory proposed by Wolfe (1997) describes essay scoring as an interaction of two cognitive frameworks: a framework of writing (i.e., content focus or interpretive framework) and a framework of scoring (decision-making processes). Figure 1 represents the interplay of the writing framework and scoring framework in the theory proposed by Wolfe (1997).

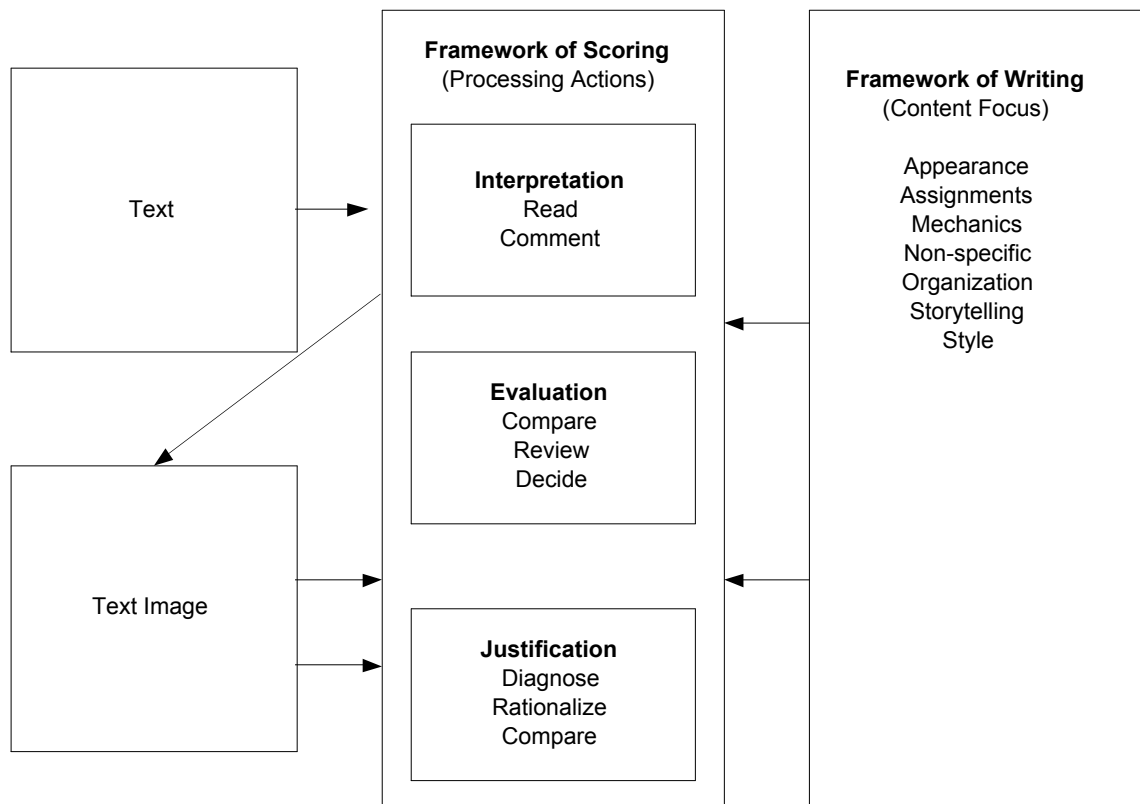


Figure 1. *Theory of reader's cognition scoring essays (from Wolfe, 1997)*

A framework of writing is a mental representation of the scoring criteria created by the judge. The most typical sources of information used for creating a framework of writing are scoring rubrics and exemplar writing samples. See, for example, the research reported by Pula and Huot (1993).

A framework of scoring is a representation created by the judge of the mental processes used to score an essay. The framework of scoring serves as a guide for how a variety of mental processing actions are used to read and evaluate the essay. In essence, a framework of scoring serves as a script that the judge uses to insure fair and accurate scoring. The framework of scoring involves creating a text image of the essay, comparing the image with the scoring criteria, and generating a scoring decision.

The representations that comprise the framework of scoring within the theory proposed by Wolfe (1997) roughly correspond to the three processes identified by Freedman and Calfee (1983).

- Process of interpretation: Representation of the processes readers use to take in information and build a text image. The reader constructs a text image of the student's composition by taking information from the printed text. The reader interprets the essay based on his/her own world knowledge, beliefs and values, reading ability, and knowledge of the writing process.
- Process of evaluation: Representation of the processes readers use to evaluate the text image, to handle discrepancies or inconsistencies in the evidence, and to weigh different features of the writing when deciding the score; and
- Process of justification: Representation of the processes readers use to monitor their own performance and to incorporate corrective feedback into their scoring activities to articulate the judgment.

Wolfe and Feltovich (1994) used this theory of reader cognition as a way of organizing the results from two studies. The first study focused on individual differences in the cognitive actions used by readers with different levels of experience scoring essays. The second study focused on the scoring methods used by readers with different levels of scoring proficiency. The authors reported that the theory proposed by Wolfe was an adequate representation of both the processes and the knowledge that are called upon by professional essay scorers.

To further test the theory proposed by Wolfe, Wolfe and Ranney (1996) observed readers' behaviors and identified eight categories of processing actions described by the theory. They used the eight categories along with the content focus elements to identify scoring procedures used by readers who were classified as "proficient," "intermediate" or "non-proficient." The proficiency of the readers was determined by high, medium and low levels of interrater agreement with the scoring rubric. The authors offered three conclusions:

- Readers focus on similar features of an essay as they formulate scoring decisions, regardless of their level of inter-rater agreement;
- Proficient readers tended to use an “interpret-then-evaluate” method of scoring whereas intermediate and non-proficient readers tended to use an “interpret-evaluate-interpret-evaluate” method that breaks the scoring task into subtasks. For example, proficient judges would read the text straight through followed by evaluation. Intermediate and non-proficient judges would read part of the text followed by evaluation, then return to reading the text followed by evaluation. Wolfe and Kao (1996) suggested that proficient judges used a linear approach to scoring whereas intermediate and non-proficient judges used an iterative approach.
- The processing actions used by proficient judges were less variable than the processing actions used by the intermediate and non-proficient judges. The proficient judges approached the task of scoring more consistently than the other two groups.

A comparison. Any comparison between the scoring process used by the IEA and theory of readers’ cognition proposed by Wolfe (1997) must note several similarities between them. First, both the IEA and human readers follow the same three process steps to arrive at a score:

- Build a text image by reading the text;
- Evaluate the text image; and,
- Articulate the evaluation.

Second, both human readers and the IEA learn from scored examples. According to empirical research, exemplar writing samples are among the most typical sources of information used for creating a framework of writing. Similarly, writing samples are used to train the IEA. Furthermore, both human readers and the IEA focus on the criteria generally included in rubrics: content, style and mechanics. For human readers, content style and mechanics are criteria generally found in scoring rubrics. Similarly, the IEA combines content, style and mechanics using multiple regression on scores assigned essays by humans as part of a training sample. The multiple-regression equation is constrained so that the content meta-variable is always assigned the greatest weight.

Finally, like an expert human reader, the IEA first creates an image and then evaluates it. According to the empirical research, proficient human readers tended to use an “interpret-then-evaluate” method of scoring whereas intermediate and non-proficient readers tended to use an “interpret-evaluate-interpret-evaluate” method that breaks the scoring task into subtasks. For example, intermediate and non-proficient readers would read part of the text followed by evaluation, then return to reading the text followed by evaluation. Similarly, the IEA creates a representation of the entire text and computes variable values for content, style and mechanics before assigning a score based on a multiple-regression equation.

In at least one aspect of scoring, human readers are unlike the IEA. Unlike human readers, the IEA does not explicitly incorporate the rubric criteria. According to empirical research, the most typical sources of information used for creating a framework of writing are scoring rubrics and exemplar writing samples. The IEA does use exemplar writing samples in training to score. However, the IEA has no process through which the criteria represented in the scoring rubric could be explicitly incorporated. Alternatively, the criteria represented in the scoring rubric are implicit in the papers used to train the IEA.

Discussion and Conclusions

This paper examined validity evidence for the IEA scores as measures of writing achievement using the framework described by Yang, et al. (2002). This framework delineates three approaches to validation studies: examine the relationship among scores given to the same essays by different scorers, examine the relationship between essay scores and external measures, and examine the scoring processes used by the IEA. This paper offered evidence from each of these three approaches.

The weakest evidence for the validity of the IEA scores as measures of writing achievement was provided by results scores given to the same essays by different scorers. All four of the measures of the relationship among scores given to the same essays by different scorers (percent agreement, Spearman rank-order correlation, kappa statistic and Pearson correlation) indicated a stronger relationship between two human readers than between the IEA and a human reader.

In contrast, the results from examining the relationship between essay scores and external measures and examining the scoring processes used by the IEA provided stronger evidence for

the validity of the IEA scores as measures of writing achievement. All four of the measures of the relationship between essay scores and expert scores (percent agreement, Spearman rank-order correlation, kappa statistic and Pearson correlation) indicated a stronger relationship between the IEA and experts than between readers and experts. In addition, the results of examining the scoring processes used by the IEA showed that the IEA used processes similar to a human scorer. Furthermore, the IEA scoring processes were more similar to processes used by proficient human scorers than to processes used by non-proficient or intermediate human scorers.

This study had several limitations. First, the design of the study obscured the comparison of the effectiveness of automated essay scoring across different grade levels. This study included three different grade levels: grade 4, grade 8 and grade 10. This study found no consistent pattern of differences in the effectiveness of automated essay scoring across different grade levels. Automated essay scoring did not appear more effective at one grade level compared to another grade level. However, differences in the kind of prompt used across grade level made conclusions difficult. For example, three different genres of writing prompt were used and the genre of the prompt was confounded with grade level. The grade 4 prompts were an expository and a narrative writing prompt. The grade 8 prompts were an expository and a persuasive writing prompt. The grade 10 prompt was an expository writing prompt.

Another limitation was the similarity between the external measure, experts' scores, and readers used in the study as different scorers. The scores assigned by experts were considered external measures because the IEA was trained using scores from readers rather than experts. Readers' scores were treated as scores given to the same essays by different scorers because a set of papers scored by readers, but not included in the analyses, was used to train the IEA. However, the difference between experts using the rubric to score essays and readers using the rubric to score essays may be perceived as a difference in degree not kind.

Despite these limitations, this study provided relatively strong, positive evidence for the use of IEA scores as measures of writing achievement. Further research with the IEA in other assessments and grade levels will be helpful in generalizing the results of this study and strengthening the validity of IEA for scoring writing assessments.

References

- Burstein, J. C., Kukich, K., Wolff, S., Lu, C., & Chodrow, M. (1998, April). Computer analysis of essays. Paper presented at the annual meeting of the National Council of Measurement in Education, San Diego, CA. (<http://www.ets.org/research/dload/ncmefinal.pdf>)
- Burstein, J. C., Marcu, D., Andreyev, S., & Chodorow, M. (2001, July). Towards automatic classification of discourse elements in essays. In Proceedings of the 39th annual meeting of the Association for Computational Linguistics and 10th Conference of the European Chapter of the Association for Computational Linguistics (pp. 90-97). San Francisco, Morgan Kaufman Publishers. (<http://www.ets.org/research/dload/burstein.pdf>)
- Clauser, B. E., Kane, M. T., Swanson, D. B. (2002). Validity issues for performance-based tests scored with computer-automated scoring systems. Applied Measurement in Education, 15(4), 413-432.
- Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The Intelligent Essay Assessor: Applications to Educational Technology. Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1(2). (<http://imej.wfu.edu/articles/1999/2/04/index.asp>).
- Freedman, S.W., & Calfee, R.C. (1983). Holistic assessment of writing: Experimental design and cognitive theory. In P. Mosenthal, L. Tamor, & S.A. Walmsley (Eds.), Research on writing: Principles and methods (pp. 75-98). New York: Longman.
- Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. Educational Measurement: Issues and Practice, 18 (4), 5-9.
- Hardy, R. (1996). Performance assessment: Examining the costs. In M. B. Kane & R. Mitchell (Eds.), Implementing performance assessment (pp. 107–117). Mahwah, NJ: Lawrence Erlbaum Associates.
- Hoover, H. D., & Bray, G. (1996, April). The research and development phase: Can a performance assessment be cost effective? Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.
- Khatti, N., & Sweet, D. (1996). Assessment reform: Promises and challenges. In M. B. Kane & R. Mitchell (Eds.), Implementing performance assessment (pp. 1–21). Mahwah, NJ: Lawrence Erlbaum Associates.

- Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to Latent Semantic Analysis. Discourse Processes, 25, 259-284.
- Landauer, T. K., Laham, D., & Foltz, P. W. (2001). Automatic essay assessment with Latent Semantic Analysis. Unpublished manuscript.
- Parkes, J. (2000). The relationship between the reliability and cost of performance assessments. Educational Policy Analysis Archives, 8 (16), 1-15.
- Reckase, M. D. (1995). Portfolio assessment: A theoretical estimate of score reliability. Educational Measurement: Issues and Practice, 14 (1), 12-14.
- Thacker, A. A., Koger, L. E., & Koger, E. M. (1998, June). Professional development under the Kentucky Education Reform Act: A study of practices in 30 middle schools (Report No. FR-WATSD-98-38). KY: Radcliff.
- White, E. M. (1986). Pitfalls in the testing of writing. In K.L. Greenberg, H.S. Wiener, & R. A. Donovan (Eds.), Writing assessment: Issues and strategies. New York: Longman.
- Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. Assessing Writing, 4, 83-106.
- Wolfe, E. W., & Feltovich, B. (1994, April). Learning to rate essays: A study of scorer cognition. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Wolfe, E. W., & Kao, C. (1996, April). The relationship between scoring procedures and focus and the reliability of direct writing assessment scores. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.
- Yang, Y., Buckendahl, C. W., Juskiewicz, P. J., & Bhola, D. S. (2002). A review of strategies for validating computer-automated scoring. Applied Measurement in Education, 15(4), 391-412.