

Evidence for Reliability, Validity and Learning Effectiveness

Introduction

Pearson's Knowledge Technologies group has conducted a large number and wide variety of reliability and validity studies for both the Intelligent Essay Assessor (IEA) and Summary Street components of WriteToLearn.

Although comparable studies have produced roughly comparable results, we focus here wherever possible on studies conducted in the last two years so as to best represent the current status of the products.

The KAT Engine. The technology underlying Summary Street, IEA, and WriteToLearn is based on the KAT engine including Pearson's unique implementation of Latent Semantic Analysis (LSA), an approach that is trained to measure the semantic similarity of words and passages by analyzing large bodies of relevant text. LSA then can closely approximate the degree of similarity of meaning of two texts as judged by human readers. This ability has been documented in a number of top-ranked refereed professional journals (e.g., Landauer & Dumais, 1997).

Summary Street

Summary Street has been evaluated in a number of language arts and subject matter classrooms. Across several smaller studies and experiments, it has proven to measurably increase students' summarization, general writing, and reading comprehension skills.

Although Summary Street and WriteToLearn cannot provide the detailed and elaborated critiques that teachers can, it does provide the opportunity for much more evaluated reading and writing practice than can the typical public school teacher—who teaches over 100 students a day. Thus, Summary Street increases the actual reading and writing that literacy researchers have shown to be the most important determiner of achievement.

University of Colorado IERI Research Results

Researchers from the University of Colorado Institute of Cognitive Science evaluated Summary Street over the course of a five-year study funded by the Interagency Education Research Initiative, a collaborative effort sponsored by the National Science Foundation, The U.S. Department of Education and the National Institutes of Health.

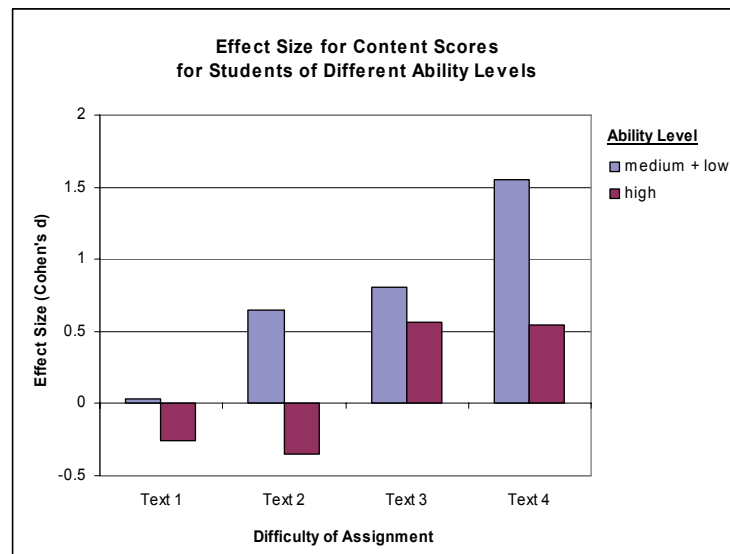
The IERI project at CU included controlled experiments comparing users and non-users of Summary Street, system measured progress in writing summaries that met content quality and ELA requirements, and data from state tests of ELA by students in field trials involving a large number of

diverse schools, including Title 1 schools, randomly assigned to be given Summary Street or not during the third and fourth year of the program.

A white paper, detailing all of the findings, entitled, “Building Student Summarization, Writing and Reading Comprehension Skills with Guided Practice and Automated Feedback: Highlights from Research at the University of Colorado” is included in Appendix C.

Here is a brief summary. In the randomized controlled experiment, students in Colorado middle schools wrote a series of four summaries of increasingly advanced readings over four days in a two week period. In some schools they used Summary Street, in others not. All summaries were blind scored with high reliability by two teachers. The experimental group students, especially those in the lower 75% of the control summary scores, showed significantly greater growth in summarizing ability.

In only four weeks of practice, the students improved their content summary scores by an overall effect size of $d = 0.9$ (see the figure below). For a class of mixed-ability students, students scoring at the fiftieth percentile improved their writing performance with difficult materials to the eighty-second percentile. When the performance of low- and medium-ability students (the lower 75 percent of the distribution) is considered, the effect size was $d = 1.5$ for the most difficult materials. (An effect size of 1.0 corresponds to approximately a one-grade difference in elementary school. An effect size of 1.5 is extremely rare in educational research.)



Effect sizes for summary content scores for low- and medium- (bottom 75 percent) vs. high-ability (top 25 percent) students. Based on 120 eighth-grade students randomly assigned to a control group with no feedback and an experimental group.

As corroborating evidence, independent scorers found increases not only in the student's ability to condense and abstract, but also in measures of organization and style. Repeated, guided practice led to improvement in

writing ability across a broad range of measures. In the analysis of the long-term data, the before and after summary performance of matched classes showed that the increase in performance endured, even when students wrote without the support of Summary Street. Students included 50 percent more relevant content in their test summaries after having used Summary Street in their school work. Finally, students who used Summary Street increased their performance on gist-level comprehension items of a standardized comprehension test (Colorado State Assessment Program or CSAP) by an effect size of $d = 0.42$, compared with students who practiced summary writing without getting feedback. Scores on CSAP reading comprehension items were 1.4 standard deviations higher for the lower 75% of students after just two weeks—four summaries—with Summary Street than were those from matched control classes.

The results of this work have been reported in refereed professional journals and conferences, and edited volumes (see Caccamise, Franzke, Eckhoff, Kintsch & Kintsch, in press; and Franzke, Kintsch, Caccamise, Johnson & Dooley, 2005).

Earlier Study Results

In the earliest classroom trials of Summary Street in 2001, two 6th grade ELA classes in a Boulder, CO middle school used either the then full-featured system or a system with the same interface and interaction features but no feedback, essentially a word processor. Students in both conditions wrote summaries by hand in their classrooms, then typed them into an input box on a classroom computer, revised them until satisfied, then handed them in to the teacher for grading and comment. The experimental group (one of whom suggested the name “Summary Street”) used the system during three class hours during the first of three weeks of the experiment, and thereafter used the no-feedback version. The other group was in the no-feedback condition the first and third week.

After the last week, the class teacher and another ELA teacher each rated all of the essays independently in random order and without knowledge of the condition in which it was written. Ratings included adequacy of summary content, style and mechanics. Summaries written with Summary Street were significantly better on all counts. Summaries written without Summary Street feedback after using Summary Street retained a significant advantage.

Students voluntarily spent twice as much time, on average going beyond class periods, revising than did those without Summary Street. Interviews produced highly favorable evaluations by teachers and students. After the experiment, the teachers requested continued use, as did some teachers who had seen or heard of Summary Street but not used it. Some of them eventually got access and used it without training.

Intelligent Essay Assessor

Evidence about the Intelligent Essay Assessor

For essay scores, the same measure is customarily used to evaluate reliability and validity, how well humans—and in this case humans and an automated scoring system—agree, expert human opinions about writing being in some respects the final arbiter of writing quality. However, Pearson believes that this criterion is subject to some amount of doubt and to important supplementation when the goal is the test's use in developmental or formative assessment. On the one hand, human inter-rater reliability may be affected by training readers to agree where uninfluenced judgment might not, and some aspects of writing, such as spelling, may actually be more consistently evaluated with computer help. On the other hand, other criteria, such as whether scores vary appropriately with known differences in test-taker abilities, such as having read or studied relevant lessons or content or being in earlier or later grades, or whether they correlate with or predict other measures of literacy knowledge and skills, or their formative use results in measurable improvements in literacy performance can be of equal or greater interest. Therefore, our various studies have included reliability and validity measures of several different kinds. They have included independent teacher judgments without scoring collaboration or instruction, administration of tests before and after studying relevant material, scoring by readers with different levels of expertise, comparing performance of students in different grades and of different chronological ages (vocabulary and other language skills improve by cultural immersion), correlation with other kinds of tests, effects on scores on other tests taken later and improvements engendered and measured by taking the test itself plural times.

In reviewing these studies, we focus on our largest, cleanest and most statistically and methodologically precise studies and experiments, but also describe some of lower power that add wider perspective on the informative value of the assessment.

Introductory methodological note: For validation, we favor product-moment correlations between continuous KAT engine scores and whatever score the human graders give over agreement on grade or score categories. This reduces to the minimum the contribution of quantification error and exploits the greater precision of our machine scores than the ~2 bit “Miller's Law” limit of human absolute judgments. It also avoids classifying the scores into discrete score groups, a matter that often involves pedagogical and policy decisions largely irrelevant to questions of measurement precision. Nonetheless, we also analyze exact, adjacent and discrepant scores so as to report results in the customary manner.

Our most recent, 2006, and elegant study was conducted in collaboration with MetaMetrics (MM) Inc. using holistically scored essay data that they had collected. We regard this study as research into the fundamental accuracy of our automatic essay grading methods. The data were 3,453 moderate length essays written by 4th, 6th, 8th, 10th, and 12th grade students on 18 different

commonly used ELA prompts in a design in which each student answered 4 prompts, each on a different day, each scored by 4 different experienced and independently working readers, with all variables factorially counterbalanced such that the effect of grade, student, prompt, day of test and reader could be measured and partialled out of the final score, which was computed by a Rasch Lexile method. With these data, the automatic scoring engine had a reliability correlation of .91 with the human readers, identical to the very high mean human-human reliability.¹

With these same purified residual data we were also able to analyze how well human and IEA scores compared in measuring differences between scores on the same essays when taken by students in different grades, thus points in normal progress in writing skills. Of the 18 prompts, three had been answered by both 4th and 6th grade students, three by 6th and 8th, three by 8th and 10th and three by 10th and 12th, each pair containing a maximum of 108 and minimum of 98 students in both grades.

The average absolute difference between the scores of students separated by two school grades had effect sizes of .45 and .43 respectively for human and IEA. These average changes were highly significant statistically, $p < 10^{-4}$ in both cases. But the difference between them was not, p one-tailed $\sim .73$, i.e. no appreciable difference.²

We had performed a similar analysis in 2002, on holistic scores from a sample of 900 minimally constrained open-ended creative fiction essays. These had reliability correlations of .90 for both human double scoring and IEA to human scoring. Student school grades ranged from 5th to 7th with varying numbers of responses, and we had chronological ages of each student as well. In that case, correlations of human and IEA sensitivity to both school grade chronological age were significantly, $p < .01$, higher for IEA. We hypothesize that the apparent superiority of IEA in that case might be attributable to greater amounts of noise in the development related component of the score variance and less in other components for the human than for the machine grading scores.

Using the purified residuals data in our 2006 study, we also explored new ways to score essays without regard to the prompt being answered using only features common to good content in any essay and on IEA measurements of the quality of writing qua writing. After some exploration, we attained .87 reliability with a variant of the usual IEA scoring model. Although this reliability will undoubtedly shrink somewhat in the noisy case of ordinary assessment, the new method adds another kind of free-response assessment to our quiver, one that will be especially useful for Summary Street in which no pre-scored training data are used.

Correlations are shown below for prompt independent scoring—scoring essays on all 18 prompts with the same scoring engine without retraining or re-calibration—for the same essays as described above.

N	Pearson Correlation	Human-human
3,453	0.87	0.80

Data on IEA Reliabilities Over Grade Levels 6 to 12 With 81 Prentice Hall Prompts

Our most extensive reliability evaluation study with operational data was conducted with results from 33, 205 essays written to 81 different Prentice Hall ELA prompts from their textbook companion website, 10 to 15 different prompts per grade level, over seven grade levels, 6th through 12th. The overall mean inter-rater reliabilities for holistic grades are given in the table below.

	Human-Human	IEA Human
Correlation	0.86	0.90
Exact agreement	61.7	61.1
Exact + adjacent agreement	97.7	98.1

IEA-Human correlations were higher than Human-Human for all seven grade levels, the difference in r ranging from .03 to .06.

Overall, IEA was correlated with averaged pairs of human scores on essay significantly better than the humans scores were correlated with each other ($p < .000$, by t-test 2-tailed paired comparison, $df = 6$, over the seven means of grade level scores. This way of estimating the statistical significance avoids inflating the df with between-prompt, between grade-level and over possible multiple scores by same student that would have occurred if the number of individual essays were taken as the basis. It is an application of the Greenhouse-Geisser approach.) The degree of superiority of IEA in this case, 4% in the product-moment correlation coefficient, 9% greater shared variance, is unusual and may be of practically important magnitude for reflecting the small incremental skill gains expected in formative developmental assessment.

Exact agreement was almost identical for human to human and IEA to Human, $p < .54$, by the same statistical method. The contrast between correlation and exact agreement for humans and IEA could be because distributions over score categories are slightly different from those of human readers.

Trait Scoring for Essays

In several trial and operational uses of the Intelligent Essay Assessor, we have also provided scores on various components or traits of English writing. Generally human scores on traits have considerably lower inter-rater reliabilities than do holistic scores, and the trait scores are all highly correlated with each other and with holistic scores. IEA scores trained and calibrated on the human scores to match them are therefore also lower and highly correlated with each other. Results are shown below for one exemplary set with a large variety of traits and better than usual double scoring reliability by humans. This IEA scoring was done by Pearson's Knowledge Technologies group for an independent large professional testing company client.

Correlations (r) among trait scores:

Description	Holistic	Content	Effective sentences	Focus & Organization	Grammar, usage, mechanics	Word Choice
H-H r between holistic and traits	0.71	0.80	0.74	0.82	0.60	0.69
IEA-IEA r between holistic and traits		0.81	0.75	0.82	0.62	0.71
H-IEA r on the same traits	0.72	0.81	0.76	0.82	0.74	0.71

Human-to-Human correlation and Human-to-IEA correlation on the holistic scores were almost identical, $r = .71$ and $.72$ respectively. The correlations of individual traits with the Holistic score were also very similar for IEA and for humans scores (the correlation between the correlations over the five individual traits was $r = .84$.) [The mean difference favoring the Human-to-IEA agreement over the Human-to-Human agreement was very small and not statistically reliable, $p < .15$.]³

Conclusions

The practical overall conclusions from the body of reliability and validity evaluation data are (a) that human and machine scoring methods are very nearly equivalent, (b) that the use of standard IEA for formative assessment and feedback for essays as used in WriteToLearn is well justified, and (c) that use of IEA for WriteToLearn feedback and for formative developmental assessment is sufficiently robust to serve a steering function for both day-to-day student learning and teacher guidance, as well as for district and state oversight for ELA learning.

Endnotes:

¹ With raw scores instead of the residuals the IEA to human correlation was $.73$.

² The small size of these differences (which were all positive except that for the 10th to 12th comparison, which was, puzzlingly, slightly negative), an average effect size gain of only $.23$ sigma per year, and a total gain of only 1.7 sigma over the whole 4th to 12th period. While it might be thought this was due to restriction of range by the 6 point scoring rubric, the fact that the analyzed scores were noise-reduced averages of 4 independent ratings on fairly large samples casts some doubt on such an explanation.

³ We note that in some scoring program that have elicited very short or highly unusual kinds of essays, or used highly unusual scoring procedures—for example, > 10 -point scores with erratic category distributions, IEA has not equaled human reliabilities. WriteToLearn, however, uses only testing and scoring procedures for which IEA has proven reliable.

