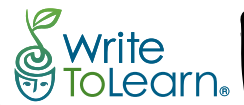# Demonstrating Reading & Writing Performance Gains
Efficacy Report

Practice for Reading & Writing

WriteToLearn® is a web-based tool designed to teach and assess reading comprehension and writing skills by making learning and practice enjoyable and rewarding. WriteToLearn combines summarization and essay writing activities. Used for both instruction and assessment, WriteToLearn gives students immediate, targeted feedback needed to develop the reading comprehension and writing skills that are critical for academic success. It provides teachers with the greatly expanded ability to assign reading and writing practice with immediate and highly useful, personalized feedback.

WriteToLearn focuses on writing for content knowledge expression across a variety of academic subjects, such as science, social studies, and history. It also contains essay prompts similar to those used for state writing assessment and college entrance exams.

Overview

First, general methods that have been proven effective in controlled classroom studies of reading and writing proficiency are reviewed. Second, data from controlled classroom and case studies of students using the components of WriteToLearn to master summary writing and essay writing are presented. Brief discussions of how and why WriteToLearn works concludes.

What works to increase reading & writing effectiveness?

The *Reading Next* (2006) and *Writing Next* (2007) reports provide recommendations for effective literacy programs based on those practices that have been shown to work across hundreds of controlled classroom studies. The most effective

teaching strategies for reading and writing are the following, with effect sizes where reported:

- Teaching students strategies for planning, revising, and editing their compositions (*Writing Next*, effect size 0.82)

- Explicitly and systematically teaching students how to summarize texts (*Writing Next*, effect size 0.82)

- Direct, explicit comprehension instruction (*Reading Next*)

- Effective instructional principles embedded in content (*Reading Next*)

WriteToLearn directly supports these important strategies for reading and writing. The product gives immediate and specific feedback that helps students practice writing through review and repeated edit and critique cycles. The summarization portion of WriteToLearn lets students practice summary writing across diverse content areas and again provides instant feedback. For summaries, the feedback captures how well the student has covered the content of each major section of the document that the student has read. The read, write, and revise cycle encourages the students to reread and re-express those parts of the text that they have not as well understood.

All WriteToLearn feedback is targeted at frequent revision by providing:

- **For essays —** An overall holistic essay score; scores and feedback on six traits (ideas, organization, conventions, sentence fluency, word choice, voice); and feedback on spelling, grammar, and repeated information

- **For summaries —** Section-by-section coverage and feedback on appropriate length, unimportant and redundant content, and copying from the text along with revision hints.

The average student revises an assignment six times; practice and more practice leads to proficiency.

WriteToLearn is Effective

Summary Street: The summarization component of WriteToLearn

Summary Street: The summarization component of WriteToLearn

Summary Street was the result of joint development between researchers at the University of Colorado and scientists and developers at Knowledge Technologies. Supported by a variety of grants from private agencies, federal Small Business Innovation Funds and a five-year Interagency Education Research Initiative (IERI) grant awarded in 2002, the efficacy of Summary Street has been thoroughly demonstrated. The studies are described in temporal order, progressing from an early pilot study to a large multi-school efficacy trial.

Study 1. Summary Street use produced higher quality summaries compared to a control group.

In the earliest field study, 60 students from two sixth grade classes in the Boulder Valley School District participated in a controlled experiment. Each student wrote two texts in counterbalanced order, one using Summary Street and the other using a standard text editor. The results showed that the students who used Summary Street:

- Received higher grades on their summaries as assessed by teachers blind to the condition to which the student was assigned

- Spent longer on the writing task

- Retained the skills they learned well after they stopped using the tool

Study 2. Students showed large benefits in summarizing with only one month's use.

Franzke, E. Kintsch, Caccamise, Johnson and Dooley (2005) had students use Summary Street in four classes for four weeks. Students improved their content summary scores by an overall effect size of d = 0.9. This means that for a class of mixed-ability students, students scoring

at the fiftieth percentile raised their writing performance with more difficult materials to the eighty-second percentile.

When the performance of low- and mediumability students (the lower 75 percent of the distribution) was considered, the effect size rose to d = 1.5 for the most difficult materials. (An effect size of 1.0 corresponds to approximately a onegrade difference, e.g. from fifth to sixth grade.)

Improvements to writing quality.

Blind scoring by teachers on the summaries produced in Study 2 were graded as superior on:

• Overall quality

• Completeness of content coverage

• Organization

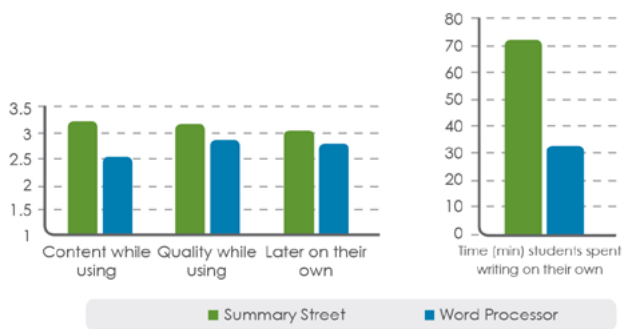• Stylistic quality

• Number of irrelevant details included

Figure 1. Summary Street produces better essays as judged by teachers in a two-week trial of sixth grade students.

Study 3. An evaluation study demonstrated that beneficial effects of summarizing are directly related to its usage.

A large two-year efficacy study was conducted by University of Colorado researchers in nine Colorado school districts with 2,851 students in grades 5-9 (see Caccamise, Snyder, Allen, DeHart, E. Kintsch, W. Kintsch, and Oliver, in preparation).

Classes of students were assigned to either use Summary Street or to receive traditional teacherprovided summarization instruction.

Of the students who used Summary Street, most of them used it for an average of 5-6 different texts during the year. Students were given both a summarization pretest at the beginning of the school year and at the end of the school year, as well as a standard short reading comprehension test (Test of Reading Comprehension, or TORC) at the beginning and end of the year.

The experimental group was superior to the control group in summary writing for both years. The improvement in summarization was highly related to the number of texts a student studied and summarized during the year, as well as the amount of time students spent using the learning tool. Comprehension improvements on the TORC test were highly related (p < .002) to the amount of Summary Street use (see Figure 2).

In conclusion, research demonstrates that WriteToLearn's Summary Street improves overall reading comprehension and both the content and style of writing. And it produces these effects in as little as one month's time. Also, students spend twice as much time writing with Summary Street than without, thus spending more time on task.
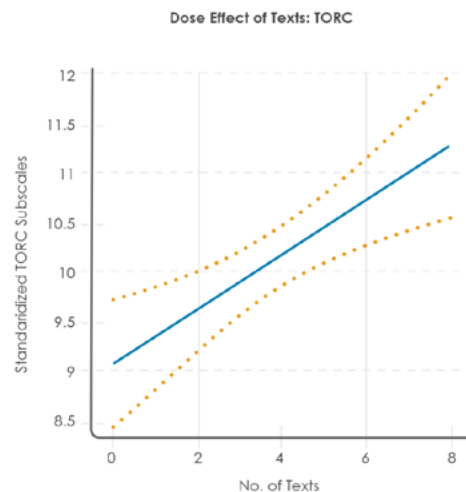

Figure 2. Performance on a standard reading comprehension test as a function of the number of texts studied with Summary Street during the school year with the 95 percent confidence interval shown in red. The pre-test scores for the same test were used as a covariate to control for student ability.

The Intelligent Essay Assessor (IEA): The essay writing component in WriteToLearn

The Intelligent Essay Assessor's performance equals or surpasses human grading.

Thoughtful teacher comments on writing assignments help students become better writers. However, the turnaround time between a student's submission and teacher feedback is measured in days or even weeks. Plus, a teacher teaching over 100 students a week cannot possibly read and comment on as many essays as the automatic scorer in the Intelligent Essay Assessor. With fewer writing assignments, students receive fewer opportunities to perfect the skill. An automated tool like WriteToLearn removes these time barriers and enables more practice. As a result, students can receive instantaneous, accurate feedback on overall essay quality, traits of writing, grammar, spelling and repeated information. At the same time, the speed and care with which teachers can evaluate a smaller number (including those also automatically scored) is increased.

Each essay prompt is scored by the Intelligent Essay Assessor (IEA) using scores assigned by human raters to several hundred representative student essays all written in response to a particular essay prompt or question for a particular grade level.

By using computational modeling, IEA mimics the way in which human readers score. In study after study comparing the performance of IEA to that of skilled human graders, the quality of IEA's assessment equals or surpasses that of the humans.

Note: The Intelligent Essay Score is also closer to a true score (i.e. more accurate), because the IEA score is based on the average of multiple graders rather than the score of one or two. In a study comparing grading of the same prompts by the college professor, the graduate student teaching assistant

and undergraduate students, the IEA correlated highest with the professor, next with the graduate students, and last with the undergraduate graders. See Figure 4.

Prompts	N	Machine-Human Reliability	Human-Human Score Reliability
Prentice Hall LA (6-12)	81	0.89	0.86
MetaMetrics	18	0.91	0.91
Higher Education Prompts	8	0.86	0.83

Figure 3. Reliabilities between human graders and the Intelligent Essay Assessor for several sets of essays: Prentice Hall Language Arts prompts for grades 6-12; a large balanced essay data set collected by MetaMetrics, and performance-based prompts for higher education.
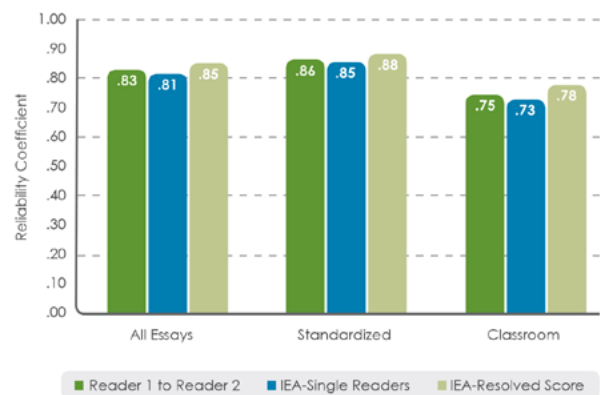

Figure 4. Reliabilities between different prompt types; those used in a standardized test and those used in a classroom application. The reliabilities are between (1) two human raters (2) IEA and a single human rater, and (3) IEA and the average of human raters.

How the Intelligent Essay Scorer works: a short essay for the mathematically inclined.

First, a set of representative student essays are collected and scored independently by two or more human graders. Usually 200 to 250 doubly scored human papers are sufficient. A regression model with about 50 content and computational linguistic variables is used to predict the average human score. A separate regression model is calculated for each essay prompt.

By far the most important variable for matching human scores turns out to be the essay's content. This variable uses Latent Semantic Analysis (LSA). Latent Semantic Analysis is a computer model that was invented and patented by several Knowledge Technologies employees in the late 1980s and is now in wide use around the world.

LSA automatically constructs a semantic space (a number representing the meaning of each word) by analyzing large volumes of text that an average student would encounter and read through high school. The text corpus for this includes all the paragraphs from about 12 million running words of text. LSA uses as input a cooccurrence matrix of words and their frequency in paragraph units. This input matrix is reduced to one of much smaller rank, using Singular Value Decomposition (SVD), a matrix algebra technique similar to factor analysis. SVD is a least squares approximation of the original matrix. It usually uses 300 independent vectors to represent each word and each paragraph in the text collection. In the end, the analysis assures that every paragraph is the sum of the 300 element vectors for its words, and every word is the average of all the vectors standing for the paragraph that uses the same vocabulary corpus, not just those already in the corpus. A variety of analyses and applications have found that LSA usually agrees with the human judgments of the similarity of two paragraphs or words 90 percent as well as two humans agree with each other.

For scoring an essay, the 200 to 250 training essays are each given a 300-dimensional score by averaging the word vectors occurring in each essay. That is, each word is represented by a vector with 300 real numbers corresponding to each of the dimensions — the separately measured quantities describing the essay. New essays to be graded are given a 300 dimensional score using the words that occur in them and averaged over each of the 300 dimensions.

Next, the new essay is compared to each of the training essays in terms of similarity (cosine of the angle between the two essays or Euclidean distance between the two). The closest neighbors to the new essay and training essays determine the content score. Essays with high scores will tend to cluster. So, a new essay close to high scoring training essays will receive a high score.

Off-topic essays can be flagged automatically because they have insufficient content similarity to the training papers.

Many other automatically (thus consistently) used variables are also used to score each essay to insure that factors not captured by LSA are not ignored. Virtually all the separate characteristics of student essays on which teachers base grades, comments and corrections influence IEA scores to approximately the same extent that they do for human scorers. This is also true of the characteristics described in the rubrics that human graders seek to follow. Measures based on the raw length of essays, sentences or paragraphs are never used. Similarly, keywords, such as ones that signal an essay's organization (e.g. "first," "in conclusion," "thus," etc.) are not given special weight. These types of variables are too highly coachable. If it were known that using them increased scores, beating an automatic essay grader would be quite simple. A separate regression model is calculated for each essay prompt.

A prompt independent grading model has also been developed that will score an arbitrary essay based only on the grade level of the student. Because the scoring engine is not trained on essays responding to a particular prompt, the scoring is based on stylistic, grammar, usage, and mechanics variables. The scoring engine has no way of factoring in the content of the essay. However, it is easier and less expensive to use the prompt independent model. While a bit of accuracy is sacrificed — a decrease of ~0.1 in the

reliability coefficient — it is easy for teachers to customize the prompts to their lesson plans. The downside of the prompt independent method is that the score uses only linguistic, stylistic, vocabulary, and mechanics variables.

Student performance using WriteToLearn's Intelligent Essay Assessor

The use of WriteToLearn has lead to substantial performance gains for students as measured by state and district reading and writing tests.

Study 1. The Iredell-Statesville, North Carolina school district began using WriteToLearn in Troutman Middle School in January 2009.

Eighty-five students used WriteToLearn and a control group of 80 did not. Of the students who used WriteToLearn, the percentage that performed at the proficient level or above went from 67 percent in the previous year to over 95 percent in 2009. Only three students in the experimental group failed to show progress. The control group remained relatively stable from year to year with 70 percent attaining proficiency or above.

The benefits the Iredell-Statesville district saw in using WriteToLearn included:

1. An increase in sentence fluency

2. An increase in the length of written responses

3. An increase in the judged quality of the writing product—adding details, organization, and style

4. Immediate feedback, allowing students to make instant revisions instead of waiting weeks for a teacher to grade drafts

5. Greater interest/motivation in the writing process

Study 2. Power usage of WriteToLearn in Cherokee County, Georgia.

Students in sixth grade language arts classes taught by one teacher made extensive use of WriteToLearn over several school years.

Each year, WriteToLearn was used for more assignments. From September 2009 to February 2010, five classes had been assigned 28 essays and summaries, resulting in 23,000 drafts with an average of six revisions per assignment. Clearly, they received very much more writing practice than most classrooms. Before adopting WriteToLearn, this teacher typically assigned three essay assignments during the school year.

The results below show sixth grade proficiency data prior to the introduction of WriteToLearn and again, after a full school year of using WriteToLearn. In May of each year, sixth graders take an eighth grade essay test, which is used as a predictive measure. They are also given Georgia's Criterion-Referenced Competency Test in language arts for the sixth grade level. The CRCT assesses students' knowledge of grammar, usage, and mechanics as well as paragraph construction knowledge of topic sentences, appropriate transitions, effective sentences, etc.

Condition	Essay Writing % Proficient on 8th Grade Test	Georgia CRCT Language Arts Test % Proficient on 6th Grade Norms
Before WriteToLearn	62%	80%
After WriteToLearn	73%	90%

Figure 5. Proficiency before and after a state writing exam.

Study 3. English Language Learner (ELL) Elementary Class achieves results.

Valverde Elementary in Denver is a 91 percent Hispanic, low income, and underperforming school. Mae Guerra described her students' annual progress when using WriteToLearn as "amazing." In her class of 22 fourth graders, many students couldn't complete a sentence at the beginning of the year. WriteToLearn boosted the confidence and the ability of the students to write well-organized, thoughtful pieces. The tool improved overall writing and reading skills.

Assessment results showed dramatic before-and after differences. By the end of the 2007-2008 school year, Guerra's students scored 91 percent proficient/advanced in reading and 95 percent in writing on the district-administered benchmark tests – an improvement of 90 percent over their results from the beginning of the school year.

Study 4. An underperforming middle school in Denver increased reading and writing scores by 60 percent after adopting WriteToLearn.

At Place Middle School, the vast majority of students come from low-income households, many dealing with language barriers and other issues. One language arts teacher led his 105 sixthgrade students to a 60 percent increase in reading and writing scores on district-level benchmark tests in just one school year. At the beginning of the school year, 51 percent of his students scored in the "unsatisfactory" category, while nine months later, only 10 percent fell in this category. In August, only four percent were considered "proficient" or above; by May, 65 percent were scoring at this level. Mr. Amos, the teacher said: "I'm not attributing all of our success to WriteToLearn, but it's definitely a huge component. It picks up where my teaching leaves off."

Why WriteToLearn Works

Writing and reading are contact sports — you cannot improve very much by watching other people read and write. So a primary reason why WriteToLearn works is that it motivates students to spend more time reading, writing, and revising.

One frequently heard comment is that students engage in the task better and more willingly when they are using WriteToLearn. The engagement comes from instantaneous feedback. Students see immediate progress and understand that they can control the learning outcome. It is also game-like in its iterative feedback.

Teachers play an indispensable role when using WriteToLearn—they can add the extras that the software cannot address, such as suggesting ways to reorganize an essay, examples that might amplify the main point, and so forth. By contrast, with a traditional classroom-assigned essay, the teacher can assign only a few essays per year, limited by the hours that need to be devoted to reading and commenting on each student draft. WriteToLearn students submitted an average of six revisions for each essay assignment. Luckily, a teacher can be the recipient of the students' best efforts and have the leisure to add the personalized human insights and comments that can bring students and teachers together in the quest for competency and excellence.
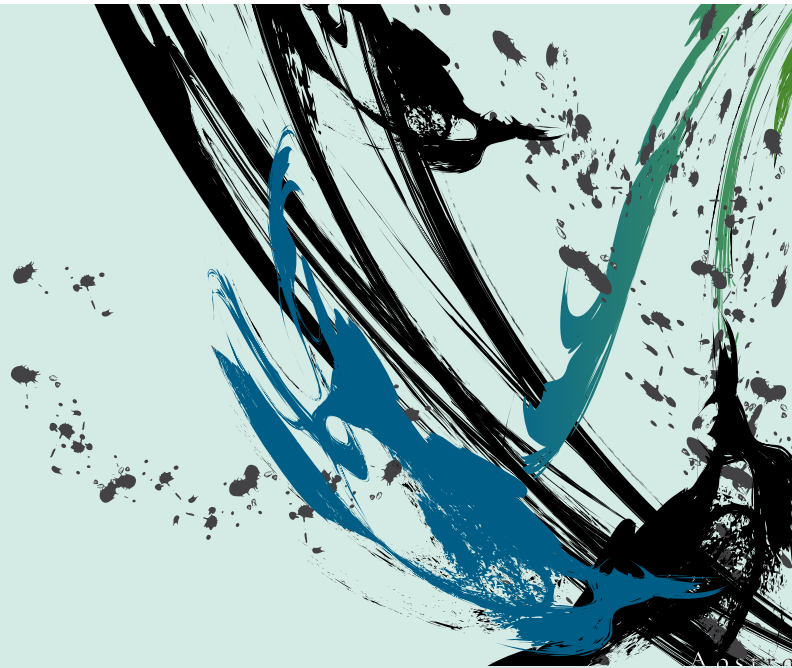
References

Biancarosa, C. & Snow, C. E. (2006). Reading next: A vision for action and research in middle and high school literacy. A report to Carnegie Corporation of New York (2nd ed). Washington, DC: Alliance for Excellent Education.

Caccamise, D., Snyder, L., Johnson, N., Allen, C., DeHart, M., Kintsch, E., et al. (in preparation). Summary Street: Scale-up and evaluation.

Franzke, M., Kintsch, E., Caccamise, D., Johnson, N., & Dooley, S. (2005). Summary Street®: Computer Support for Comprehension and Writing. Journal of Educational Computing Research, 33, 53-80.

Graham, S. & Perin, D. (2007). Writing next: Effective strategies to improve writing of adolescents in middle and high schools. A report to Carnegie Corporation of New York. Washington, DC: Alliance for Excellent Education.

Kintsch, E., Caccamise, D., Franzke, M., Johnson, N., & Dooley, S. (2007). Summary Street: Computer-guided summary writing. In T. K. Landauer, D. McNamara, S. Dennis & W. Kintsch (Eds.), Handbook of Latent Semantic Analysis (pp. 263-278). Mahwah, NJ: Erlbaum.

TORC, Test of Reading Comprehension, Pearson Assessments, Brown, V. L., Wiederholt, J. L., and Hammill, D.

Pearson