A New Formative Assessment Technology for Reading and Writing

Advances in assessment technologies are affording teachers and students new ways to efficiently assess and track achievement while also better promoting learning. WriteToLearn is one such technology, a Web-based tool that integrates practice and assessment in reading comprehension with writing about what is learned. Based on the principle of immediate feedback, WriteToLearn is a combination of summative and formative assessment tools that seeks to encourage, instruct, and reward progress in reading and writing while it is happening. It does this by providing students with instant, computer-generated evaluations of the substantive content and expository quality of writing about what they are learning. Its real-time and long-term reports about student activity and progress give teachers and schools rich information for guiding classroom instruction and curricular decisions. This article discusses the motivation for and design of WriteToLearn, as well as studies of its accuracy, reliability, and effectiveness in the classroom.
testing methods that measure and predict a multitude of aspects of knowledge and skill, and offer a vast array of practical methods for school and classroom use. Nevertheless, educators, parents, students, employers, social critics, the public at large—and most measurement professionals themselves—harbor misgivings about the limitations of current techniques. Broadly speaking, this unease has three sources. One is lack of face validity, or the difference between what students are asked to do on tests and what they do in real life. For example, choosing a best synonym on a multiple choice vocabulary test item is quite unlike how words are used in reading and writing on a day-to-day basis. Second is the extremely sparse sampling by a test of what the student should know. This worry takes expression in objections to standardized testing programs that are seen as promoting narrow teaching to the test. Third, traditional testing is seldom of direct benefit to learning. Although traditional tests can motivate students to pay attention and study, they rarely involve doing the actual tasks whose learning they assess (although there are exceptions, some math and essay tests being examples). Similarly, although good multiple-choice items and statistical sampling are sufficient to predict how much students know, and thus inform teachers about areas that need attention, they do not, by themselves, teach things not already known.

Many of these limitations do not apply to teacher designed and evaluated tests, for example, essay writing and summarization assignments, classroom quizzes, textbook exercises, and portfolios. But teacher designed tests nevertheless retain some of the limitations of large-scale tests, most notably limited teacher time for individualized pedagogical feedback.

Technologies that exploit the power of computers and the Internet can fill in many of the gaps. A familiar early example is adaptive testing, in which the next question asked depends on answers to previous questions, thus improving individual knowledge coverage. There are many other kinds of assessment functions that computational automation makes either newly possible, easier to implement, or more effective, for example, embedding assessment in ongoing learning activities, measuring how well students can produce, rather than chose, right answers, giving evaluative and instructive feedback immediately, automatically choosing the best next assignments and assessments, and providing richer information to guide teacher interventions.

The need for new assessment technologies in reading and writing is motivated by a combination of established principles of literacy instruction and the practical problems of instantiating them in traditional educational environments. In particular: (a) most students have far too few opportunities to practice reading and writing skills with feedback (Black & Wiliam, 1998); (b) students benefit most from specific, immediate, and individualized feedback that addresses content as well as surface-level features (e.g., Graham & Harris, 2005); (c) learning to read and write well comes largely from doing reading and writing, the more the better; and (d) embedding assessments in natural performance makes it better able to promote abilities that will be used in and out of the classroom. We will discuss how one assessment technology in particular, WriteToLearn, instantiates these principles.

**WriteToLearn**

WriteToLearn is a Web-based tool that integrates practice and assessment in substantive reading comprehension with expository writing about what is learned. Based on established principles of immediate informative feedback, mastery, and individualization, WriteToLearn is meant to be used frequently as a tool in which assessment is part of an integrated experience that simultaneously encourages, enhances, assesses, and rewards progress.

WriteToLearn (WTL) has two components: Summary Street, in which students read and summarize short articles, book sections, or excerpts of leveled readers, and the Intelligent Essay Assessor™ (IEA), in which students write topic-prompted essays. Both feature frequent tutorial feedback with opportunities for revision and a common graphical scoreboard with immediate display of individual progress from one revision
Figure 1. Summary Street student feedback screen. Bars show content coverage from Poor to Excellent for each section of the reading. Clicking on the section name will retrieve the section for review. Small, inverted triangles indicate content coverage on the previous revision and thus allow the student and teacher to track progress from one revision to the next. Editing tools show other quality feedback and offer various kinds of tutorial help on copying from the text, spelling, redundancy, and irrelevancy.

to the next, together with results for all activities so far completed. Figure 1 shows a student scoreboard for Summary Street; Figure 2 shows a student scoreboard for IEA.

**Summary Street**

Summary writing is one of the most widely used and best proven tools for literacy learning (Shanahan, 2005). It is not only an effective strategy for increasing comprehension while in the act of reading, but helps build lasting ease in understanding and writing well and with pleasure. Summary Street is designed to give learners much more practice—with very frequent assessment and tutorial feedback—on both reading comprehension and expository writing. Summary Street uses Latent Semantic Analysis (LSA), a form of artificial intelligence that measures the semantic similarity of one passage—sentence, paragraph, or essay—to another independent of the literal words used (Landauer & Dumais, 1997; Landauer, Laham, & Foltz, 2003). In Summary Street, LSA is used to compare the semantic similarity of students’ summaries to the readings they are summarizing. As such, it
provides a measure of how well students have understood the content of what they have read and can summarize that content in their own words.

Students initiate interaction with Summary Street by logging in using a Web browser. They select a text to summarize from the library of hundreds of available materials—most often sections or chapters from nonfiction books—enter their summaries through a simple editing window, click submit, and receive immediate feedback. As shown in Figure 1, Summary Street provides feedback on: (a) content: how well the summary covers the gist and key points of each section of the reading, (b) length: whether the summary has been adequately condensed from the original text, (c) copying: whether students have copied too much from the original text, (d) spelling: which words may be misspelled, (e) redundancy: whether there are repetitive sentences that could be combined, and (f) irrelevance: whether there are unrelated sentences that could be omitted. With each subsequent revision of their summaries, students receive additional feedback and can track progress toward goals specified by their teacher.
The Intelligent Essay Assessor (IEA)

The Intelligent Essay Assessor is intended to be used either in an integrated reading comprehension and writing practice environment, or as a second opinion on essay exams administered online, including high stakes exams. In essay scoring, LSA makes it possible for IEA to mimic the way in which human readers compare the contents of one essay with others to judge how much a particular student’s essay has accomplished—except that IEA can compare each essay with hundreds of others that have been expertly scored! (It does not try to evaluate highly unique essays, instead flagging them for human attention.) The way this works is that the computer is first fed several hundred student essays, written to the same prompt, that have each been scored by two experienced scorers. It combines LSA with other computer modeling methods to measure the degree to which an essay shares the characteristics that made human readers think an essay deserved a particular score (Landauer et al., 2003). Like other leading automated essay scoring systems, for example ETS’s e-rater (Attali & Burstein, 2006), IEA measures both holistic quality and writing traits such as content development, effective use of sentences, focus, grammar usage, mechanics, and word choice, along with more specialized measures such as semantic coherence, voice, or the reading difficulty of the essay. Due in large part to the nature of LSA and its semantic representation, IEA is especially good at matching human readers in evaluating the conveyance of substantive content (Foltz, Gilliam, & Kendall, 2000). Figure 2 shows how IEA is used in WriteToLearn to provide feedback on student writing for review and revision.

The Combination of Summary Street and IEA in WriteToLearn

Teachers log into WriteToLearn to set up class rosters, administer assignments, and monitor activity and progress. Student proficiency levels can be addressed by assigning materials at different difficulty levels, by modifying scoring thresholds and expected response lengths, and by selecting different spell-checking and grammar options. When students log into WriteToLearn, they see the list of activities—readings or essay topics—that their teacher has assigned. The student then writes a first draft of the assigned summary or essay, clicks submit, and receives feedback. The student reviews the feedback, revises his or her writing, and resubmits for more feedback, typically repeating this review, writing, and feedback cycle for three to eight revisions.

Teachers can view, in real time, automated, largely graphical reports and visualizations that display individual student and class activity and progress. Figure 3 includes a teacher overview report showing class performance on an essay writing activity. Teachers also have immediate access to all successive revisions of any student’s summaries and essays and their automated content comprehension and writing quality feedback. The cumulated results of student and class activities and progress—what was read, score means and variation, criteria attainment—on content comprehension, knowledge expression, and writing traits constitute an unusually rich source of face-valid formative and summative assessments for literacy skills. In our view, this gives ELA and subject matter teachers both a powerful new tool for intervening at the right times with the right help, and progress measurements where they can be unusually sensitive and instructionally helpful.

Research Results

How this all comes together is best appreciated by actually using or watching students use WriteToLearn. Indeed, one of its most gratifying characteristics is that most students enjoy using it. They act as if—and often volunteer that—its game-like try-and-try-again with improvement feedback is engaging and fun. In one early middle school classroom trial, instead of working for the assigned half hour with Summary Street, most students voluntarily worked through the bell! The quotes below from Summary Street students and their teachers illustrate the power of the automated technology.
It’s like a teacher that helps you get better but doesn’t give you a bad grade. (6th grade student)

... told me what I was doing wrong, what I normally wouldn’t have known. (8th grade student)

... is a mind expander... it definitely makes you think. (11th grade student)

Kids really learn to pick out the important things in the text. (Middle school language arts teacher)

The best part is that feedback is immediate, and I don’t have to read 120 summaries. (Middle school language arts teacher)

In use, teachers are almost always pleased with the ease and effectiveness of school use of WTL. Indeed, although both online and face-to-face professional training on WTL are available, teachers usually require little special preparation, often using it on their own, having heard about it only from other teachers.

Research on Summary Street Reliability, Validity, and Relation to Achievement

Evidence of the reliability and validity of Summary Street comes from several sources. Evidence of construct validity includes studies that show that older students, on average, write significantly better summaries of the same readings as measured by Summary Street. External validity is supported by results that indicate that students who used Summary Street in randomized efficacy trials scored higher on some state reading comprehension items, with performance being positively correlated with frequency of using the tool. Although the reliability of the scores provided to students by Summary Street has received limited direct evaluation,
there is little doubt that reliability and validity is sufficient for Summary Street’s motivational and learning-promotion goals (Franzke, Kintsch, Caccamise, Johnson, & Dooley, 2005; Wade-Stein & Kintsch, 2004).

**Effectiveness.** There have been several small-scale effectiveness studies and one large efficacy experiment with Summary Street. In the earliest published classroom studies of Summary Street (Wade-Stein & Kintsch, 2004), two 6th grade ELA classes in a Boulder, Colorado, middle school used either Summary Street or a system with the same interface and interaction features but no feedback. Students in both conditions wrote summaries by hand in their classrooms, then typed them into a classroom computer, revised them until satisfied, then handed them in to the teacher for grading and comment. Independent blind ratings by two teachers showed significantly better comprehension and writing progress over as little as 2 weeks of use as compared to students in the control group. The students in the Summary Street group also maintained their advantage on their later summaries in which they did not receive any automated feedback. After the experiment, the teachers requested continued use.

Another study involved four 8th grade classrooms, where half of the students in each class were assigned to Summary Street and half to a control condition (Franzke et al., 2005). Controls summarized the same materials with a common word processor without feedback, and the two groups were given the same amount of time to revise. Over a 4-week period (six 90-min sessions) the average of experimental group content scores increased by a highly significant effect size of \( d = 0.9 \) relative to controls, improving from the 50th to the 82nd percentile level. Relative improvement by the lower two-thirds of the experimental group students when summarizing the most difficult reading was greater by an effect size of 1.5, more than a grade level. The independent scorers found significantly greater increases not only in ability to condense and abstract, but also in organization and style.

**Efficacy.** Researchers from the University of Colorado Institute of Cognitive Science and Knowledge Analysis Technologies evaluated Summary Street during a five-year study in a broad sample of language arts and subject matter classrooms. While the full report of the intervention has not appeared at this writing, some preliminary results are of interest. Before-to-after performance of matched classes in which students wrote summaries with and without Summary Street found about 50% more relevant content in the treatment group. Users also increased performance on gist-level comprehension items of a standardized comprehension test (Colorado Student Assessment Program or CSAP) by an effect size of \( d = 0.42 \), compared with students who practiced summary writing without getting feedback (Caccamise, Franzke, Eckhoff, Kintsch & Kintsch, 2007; Franzke et al., 2005). An initial analysis of reading comprehension items found that the more students had used Summary Street during the trial, the more their comprehension scores improved (Caccamise et al., in preparation).

**Research on Intelligent Essay Assessor Reliability**

Essay exams are widely used because of their face validity as extended, constructed response assessments that tap students’ ability to express and use knowledge. The goal of IEA is to make it possible for students to engage in this activity more often with more rapid assessment and feedback and in more contexts, including both language arts and subject matter classes. To a great extent, reliability and validity are the same for essay tests in that agreement with human judges is the criterion. Although evaluations by us and other providers show that students’ essays get better using such systems, we know of no studies that have tested whether the tool increases other desirable outcomes, such as increasing learning, generalization, or motivation as widely assumed in pedagogical practice. But we have thoroughly established its reliability, without which promoting these other goals would be very unlikely.
Studies comparing the performance of IEA against human scorers have shown that IEA agrees with individual human scorers as well as they agree with each other (Foltz et al., 2000; Landauer et al., 2003). The most recent and thorough study used essay data that had been human scored. This study provided an unusually accurate measurement of the fundamental accuracy of the automatic essay grading method. The data were 3,453 moderate length essays written by 4th, 6th, 8th, 10th, and 12th grade students to 18 different commonly used ELA prompts. Each student answered six prompts, each on a different day, and each was scored by four different experienced and independently working readers, with all variables counterbalanced such that the effect of grade, student, prompt, day of test, and reader could be measured and statistically held constant in the final score. With these data, the automatic scoring engine had a reliability correlation of .91 with the human readers, identical to the unusually high mean reliability obtained for human scores.

Although human agreement is usually the validity criterion for essays, we added sensitivity to student progress as a confirming construct validity measure. With the same data we were able to compare differences between scores on the same essays when taken by students in different grades. 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 having at least 98 students. The average difference between the scores of students separated by two school grades were almost identical for the humans and IEA, and both were highly significant statistically ($p < 10^{-4}$).

In another large study, we used essays on 81 different reading-related topics written by students in grades 6 through 12 online in a Web-based companion to a Prentice Hall reading series. The correlation between IEA and humans was better than that between the two human readers for every grade level, a difference with probability less than one in a thousand. Exact agreements on the six-point scale were nearly identical: 61.1% exact agreement for IEA to human and 61.7% for human to human. Taken together, these results show that the automated essay test can be used effectively alone to assess student writing in grades 4 through 12, and to help track improvement in writing and subject matter knowledge with age and schooling.

Conclusion

Technologies like WriteToLearn can give teachers the ability to assess, instruct, direct, and respond to the learning needs of each student with greater speed, frequency, focus, and flexibility. WriteToLearn gives students more opportunities to practice reading and writing and to receive immediate, specific feedback on their comprehension and writing than is typically possible in a classroom setting. What’s more, WriteToLearn has been shown to score as accurately as skilled human graders and to improve the assessment scores of students who use it.

With WriteToLearn, a teacher can interact with a whole class in much the same way that a symphony conductor rehearses an orchestra or a coach practices a team, being able to know what and how each member is doing at any moment and intervene quickly in a highly personalized manner. The teacher can arrange what each student will do to learn, but the system lets every student get the feedback he or she needs when he or she needs it, even if every student needs something different at the same moment. In WriteToLearn and systems like it, status and progress can be assessed more quickly and in more use-relevant contexts than before. More important, these technologies can be used to measure and help learning from moment to moment, not just to find out later how much has been learned. Such systems can also automatically provide detailed and well-analyzed reports on data collected over time and by student, activity, and class—information that is highly useful for long-range guidance of classroom, school, and curricular decisions but virtually impossible for teachers to provide.

We believe that WriteToLearn provides a good example of applying new assessment technology in education. Much traditional instructional assessment involves relatively prolonged learn-
Classroom Assessment

ing of connected bodies of knowledge that can fruitfully, and often best, be assessed as wholes after a sufficient interval has passed to determine how well what was taught is being retained. By contrast, others—learning to read and write and mathematical problem-solving are prime examples—involves dynamic activities where errors and successes happen almost continuously and probably should, at least sometimes, be assessed very quickly and frequently for maximum benefit. Mature, interactive, computer delivered assessment can do that.

Notes

1. Summary Street technology originated in academic research and development by Darrell Laham, Tom Landauer, Peter Foltz, and Walter Kintsch at the University of Colorado and New Mexico State University in the late 1990s with funding from DARPA and the MacDonnell Foundation. Iterative design, development, testing, response scoring, analysis, and investigative applications of the technology have since been implemented by Knowledge Analysis Technologies, now the Knowledge Technologies Group of Pearson, under PIs Tom Landauer, Peter Foltz, Scott Dooley, and Karen Lochbaum.

2. One teacher’s experience with WriteToLearn is chronicled at http://www.pearsoned.com/ednews/may08pages/writetolearn.htm

3. Effectiveness is evidence evaluated in randomized experiments that carefully control how the intervention is applied. Efficacy compares results when some randomly assigned groups are given access to the intervention and others are not, under natural school use circumstances.

References


