9

COMPUTERIZED WRITING ASSESSMENT

Community College Faculty Find Reasons to Say "Not Yet"

William W. Ziegler

Community colleges exist to provide educational opportunities to a fluid population, many of whom encounter sudden changes in their work, family lives, and financial situations. For this reason, community colleges often admit, place, and register a student for classes all within a little more than twenty-four hours. This need to respond promptly explains why Virginia's community college administrators took notice when the computerized COMPASS placement test appeared in 1995. The test, published by ACT Inc., promised to gather demographic data as well as provide nearly instant placement recommendations in mathematics, reading, and writing. Several colleges piloted the test independently, and by 2000 the Virginia Community College System required all its colleges to use the test, with system-developed cutoff scores, unless they could show that other measures were superior. The Virginia Community College System now had a test that not only reported scores quickly and recorded data for easy manipulation but could be used uniformly at each college, unlike the previous patchwork of commercially published and homegrown tests used by the system's twenty-three member institutions.

Faculty had little difficulty accepting COMPASS/ESL (renamed when English as a second language tests were added in 2000) as a test for mathematics and reading once pilots had shown that it produced valid placements. Writing was a different case. The COMPASS/ESL writing-placement test is a multiple-choice editing test, requiring students to detect errors and evaluate coherence and organization within short passages. However, for most English faculty, the only valid test of writing competence is writing. Pilot testing led faculty at J. Sargeant Reynolds Community College and elsewhere in the Virginia Community College System to conclude that the COMPASS/ESL writing-placement test could not identify underprepared writers as accurately as trained faculty could by evaluating impromptu writing samples. Therefore, several colleges continued to use writing samples for placement in composition, exempting

only students with high standardized test scores, which had proven to correlate strongly with the ability to produce satisfactory writing.

Unfortunately, trained faculty raters need time to read, so students and counselors facing registration deadlines still fidgeted daily while waiting for faculty to evaluate dozens of writing samples. Faculty did not always enjoy the process, but they disliked even more the idea of giving up a direct writing measure in favor of a grammar test alone.

So, when ACT Inc. introduced e-Write as a component of the COMPASS/ESL test, faculty and administrators again took notice. The e-Write test elicits writing samples using a set of argumentative prompts. Each prompt describes a simple rhetorical context, such as a letter intended to influence government or educational leaders to make a policy decision. The college can designate a time limit or allow untimed testing. Test takers type their samples in a bare-bones word processor—not much more than a message window—and submit it via the Internet to ACT for electronic evaluation by the IntelliMetric Essay Scoring Engine, which returns an overall placement score on an 8-point scale as well as five analytic scores, each on a 4-point scale. The placement score, while not as prompt as the multiple-choice writing test, arrives in a few minutes. For those thinking of placement as a customer-service function, here was the answer: analysis and direct assessment, plus the advantages that had attracted them to COMPASS/ESL at the beginning—speed, accessible data, and uniform placement practices across the state system.

Most English faculty in Virginia's community colleges would probably agree with Joanne Drechsel's (1999) objections to computerized evaluation of writing: it dehumanizes the writing situation, discounts the complexity of written communication, and tells student writers that their voice does not deserve a human audience. However, faculty at the two colleges conducting pilot studies of e-Write (Tidewater Community College and J. Sargeant Reynolds Community College) did not object to the trial. Some may have been mollified by the prospect of serving students more quickly, others by the wish to be rid of a burden. And others may have reasoned cynically that for students used to receiving a reductive, algorithmic response to their writing (put exactly five sentences in each paragraph; never use *I*), one more such experience would not be fatal.

As it turned out, composition faculty never had to fight a battle for humanistic values on theoretical grounds because the pilot studies showed e-Write could not produce valid writing placements. Among the findings:

- Both the overall scores and the five analytic scores tended to cluster in a midrange. Few samples received scores other than 2 or 3 in the five 4-point analytic scales, while 82 percent received overall scores of 5 or 6 on the 8-point scale. No scores (other than the few at the extremes of the scales) corresponded closely with instructors' ratings of the samples.
- A follow-up survey of students' grades showed that e-Write scores did no better at predicting success in the college composition course than faculty reader scores.
- The IntelliMetric Essay Scoring Engine was at a loss more often than hoped. When the artificial-intelligence engine cannot score a sample, the writing is evaluated by human raters at ACT at a higher cost and after a longer time—a day rather than minutes. More than 25 percent of the pilot samples stumped the scoring engine and required human assessment.

THE E-WRITE PILOT AT J. SARGEANT REYNOLDS COMMUNITY COLLEGE

For one week in July 2003 the college suspended its normal writing-placement process, in which students complete the COMPASS/ESL writing (grammar) test followed by a writing sample for those whose grammar scores fall below the 65th percentile. Instead, students were asked to complete the COMPASS/ESL e-Write test. Those who preferred not to use the computer were offered the regular placement writing sample, which students write by hand. Forty-six students chose the e-Write option.

The e-Write pilot used three of the five prompts provided; two would not have been suitable because their fictional contexts presented situations that would arise only at a residential college. Full-time faculty members evaluated the e-Write essays, judging the writers as either developmental (assigned to ENG 01, Preparation for College Writing I) or ready for first-semester college composition (ENG 111, College Composition I). Three faculty readers took part at first. One of these evaluated all forty-six essays, a second evaluated thirty-seven, and a third evaluated twelve. Because the e-Write prompts were rhetorically similar to our own placement prompts and because we did not want to require more time from students for additional testing, we decided to use the resulting samples for actual placement if we found them suitable. However, we would rely on our own evaluations rather than the e-Write scores.

A few months later, when all English faculty had returned for fall semester, six more readers evaluated the same samples, each reading a set of twelve. Eventually, all samples had evaluations from at least two readers. Twenty-seven essays received evaluations from three readers, and ten essays had evaluations from four readers. All readers examined only the writing samples, which included the names of the student authors. Other information, such as COMPASS/ESL reading-placement scores, time spent on the test, students' first language, and demographic data, were withheld.

All faculty readers teach primarily first-year composition. In addition, four teach at least one section of developmental writing each academic year, and three are qualified to teach developmental reading, although only one does so regularly. The readers included the English program chairperson, the coordinator of developmental English, and the head of the academic division of arts, humanities, and social sciences.

Time

E-Write provides a choice of time limits; however, the Reynolds pilot used the untimed mode. We reasoned that most students would probably not exceed the one-hour limit we place on our current writing sample. In addition, an untimed mode accommodates students with special needs recognized under the Americans with Disabilities Act.

We were correct about how much time students would use to write. E-Write data showed that the mean average time spent on the test was 31.6 minutes (excluding two sessions with recorded times of over four hours, likely the result of machine error). Only five students took more than 50 minutes; times ranged from 5 to 79 minutes.

E-Write Test Scores

One limitation we observed in e-Write was its tendency to assign mid-range scores to nearly all the samples. Only three samples received overall ratings of 7 or 8; only five received ratings of 3 or 4. E-Write awarded scores of 6 to eighteen (39.1 percent) of the samples and scores of 5 to twenty samples (43.5 percent). The five analytic scores also grouped in the midrange. Out of 230 analytic scores (five scores for each of the forty-six samples), only eight scores were 1 or 4. No essay received the highest score of 4 in conventions, organization, or style; no essay received the lowest score of 1 in conventions or style. E-Write awarded a score of 3 for focus to 69.6 percent of the samples and a score of 3 in content to 56.6 percent. Ratings for style and organization were split

evenly: 50 percent each for style scores of 2 and 3; 47.8 percent each for organization scores of 2 and 3. E-Write awarded a score of 2 in conventions to 63 percent of samples.

An uneven distribution of scores is not fatal if a test needs only to facilitate a two-level placement decision. Writing instructors might not object to a test that gives scores of only 5 or 6 if they perceived consistent, relevant distinctions between each group of samples. However, we found that e-Write's overall and analytic scales could not do this to the satisfaction of our faculty.

Time and E-Write Test Scores

The samples receiving overall scores of 5 and 6, the two largest contingents, differed little in average time spent on the test: 33.2 minutes for samples scored 6 versus 32.6 minutes for samples scored 5. Not surprisingly, the few essays rated below 5 recorded a shorter average time (21.4 minutes), but two of these spent 31 and 35 minutes on the test, not much different from those with higher scores. Likewise, the few essays receiving top scores (one with overall 8, two with overall 7) were written in 36, 35, and 48 minutes—not drastically different from average times for the midrange scores.

In the analytic measures, writers who took more time enjoyed an advantage in only two areas: content and style. Samples rated 3 in content averaged 34.4 minutes, while samples rated 2 averaged 26.5 minutes. On the style scale, samples rated 3 averaged 33.8 minutes, compared to 29.4 minutes for samples rated 2. On the other scales (focus, organization, and conventions), the average time for samples receiving scores of 3 was slightly lower than for those scored 2; the largest difference was only 2.2 minutes.

The time difference between higher and lower content scores is unsurprising; presumably quantity affects the content score, although the ACT COMPASS scoring guide (2003) states distinctions in both quantitative (number of supporting reasons offered) and qualitative terms (elaboration, selection of examples, and clarity) (61). It is unclear why those receiving scores of 3 for focus, organization, and conventions required slightly less time than those receiving 2.

Reading Scores and E-Write Test Scores

The higher a student's overall e-Write score, the higher the score in the COMPASS/ESL reading-placement test was likely to be. All three students with e-Write overall scores of 7 or 8 scored in the 85th to 99th

percentile in reading (relative to all students tested at Virginia's two-year colleges). Students with overall e-Write scores of 6 averaged 86.1 (54th percentile) in reading; those with e-Write scores of 5 averaged 76.8 (28th percentile). The five writers who received 3 or 4 in e-Write averaged 64.4 (12th percentile), none scoring higher than 79 (34th percentile).

The same was true for analytic scores. The largest difference was in the focus scale, where students whose samples received 3 averaged 86.1 in reading, compared to 65.4 (13th percentile) for students with 2-rated samples. Differences were smaller in the ratings for content, organization, style, and conventions, where students with 3-rated samples averaged from 84.7 to 87.6 (47th to 57th percentile) in reading, compared to students with 2-rated samples, whose reading scores averaged from 74.8 to 77.6 (25th to 30th percentile).

Faculty Evaluations

Faculty readers were unanimous in their ratings for twenty-two of the forty-six samples, judging nineteen as composition-ready and three as developmental. Of the remaining samples, thirteen were rated composition-ready and three developmental by a split vote. Four samples received a 50-50 split vote from an even number of readers. The votes of the three July readers determined the students' formal placements. By this method, thirty-six students were placed in ENG 111 and ten in ENG 01.

To examine interrater agreement, we examined paired readers. (For example, three readers for an essay amounted to three pairs: readers A and B, readers A and C, and readers B and C.) By this method, there were 136 pairs of readers, with 64.6 percent agreeing on either an ENG 111 or an ENG 01 placement.

Faculty Evaluations and E-Write Test Score.

Faculty tended to favor samples with higher e-Write scores, but not to a degree that justified setting an e-Write criterion for placement. Nine of the eighteen samples receiving e-Write overall ratings of 6 elicited unanimous recommendations for ENG 111, and another six elicited split decisions, with ENG 111 votes predominating. Only one sample with an e-Write score of 6 received a split ENG 01 recommendation. Five of the twenty samples receiving overall scores of 5 elicited unanimous ENG 111 recommendations, and six received ENG 111 recommendations on a split vote. Four of the 5-rated samples received unanimous ENG 01 recommendations, with another three receiving ENG 01 recommendations on a split decision.

The same pattern showed in the analytic scores: a 3 score nearly always coincided with unanimous or split decisions in favor of ENG 111, whereas scores of 2 coincided with an array of outcomes, leaning more toward ENG 01 decisions. One of the most curious outcomes involved the conventions scale, where twenty-seven students received e-Write ratings of 2; of these, eight received unanimous ENG 111 ratings from faculty, while nine received unanimous ENG 01 recommendations.

Clearly, e-Write scores did not coincide closely enough with faculty judgments to persuade instructors to turn their placement function over to the test. However, what if e-Write knew better than we did whether a writing sample showed readiness for a college composition course? To answer this question, we recorded the pilot students' final grades in ENG 111 classes during the subsequent academic year.

Faculty Judgments, E-Write Test Scores, and Success in College Composition

Typically, a large but unknown number of students who take placement tests at Reynolds do not enroll in classes during the subsequent semester. Of the forty-six students who wrote e-Write samples, six did not enroll in any class at the college during the following academic year, and another nine did not enroll in ENG 111 or ENG 01 classes, although one enrolled in an ESL composition class, one completed first-semester college composition at another community college, and one transferred credit for first-semester composition from a four-year college.

Subtracting noncompleters from the pilot group leaves a sample too small to bear up under statistical scrutiny, but these students' success rates in ENG 111 are distressingly and uncharacteristically low. During the last several years, the success rate in ENG 111 at Reynolds (the proportion of students earning grades of A, B, or C) has ranged from 65 to 69 percent. The balance includes students who withdraw or earn incomplete grades as well as those who earn D or F. But in the e-Write pilot group, just under half of students who enrolled in ENG 111 completed it with grades of C or better. The two largest groups by overall score (6 and 5) differed little from each other. Six of the thirteen students whose samples were rated 6 and who enrolled in the course completed it successfully, while five of the eleven enrolling students with samples rated 5 did so. However, composition grades for students scoring 6 included more As and Bs, while C was the most common grade for those scoring 5. No students scoring at the extremes of the overall e-Write scale—3, 4, 7, and 8—enrolled in ENG 111 at Reynolds, although one (with a 7 score) completed the course elsewhere.

In only one of the five analytic scales—content—did a relatively high score tend to mark successful composition students. Thirteen of the eighteen students with content scores of 3 received grades of C or better, with As and Bs predominating. The analytic score for conventions did the poorest job of picking out successful students: three of the ten enrolled students scoring 3 on this scale completed the course successfully, compared to eight of the eleven enrolled students scoring 2.

Faculty proved no more prescient than e-Write. Half of the students whose samples drew either unanimous of split-decision ENG 111 recommendations from faculty readers completed the course successfully. Successful students receiving unanimous recommendations earned mostly A and B grades, while those receiving split decisions received mostly Cs, but the proportion of unsuccessful students was the same—about 50 percent—for each group.

Scoring Engine at a Loss—Humans to the Rescue

One feature of e-Write is the human backup. As explained in the COMPASS/ESL technical manual, "COMPASS e-Write does not score responses that deviate significantly from the patterns observed in the original training papers" (ACT 2003, 62). The choice of modal verb—*does* not rather than *can*not—suggests disdain, similar to distaste toward washing windows, but the manual explains that the rating engine has trouble with samples that are "off topic" or too brief. If the scoring engine cannot determine a rating for a sample, ACT's human readers take over, evaluating the sample on the same scoring scales and returning the results in two days. The scoring engine needed human backup in twelve of the forty-six samples (26 percent). In a typical semester, when Reynolds tests more than two thousand students, this projects to more than five hundred students for whom one advantage of e-Write—speedy response—would vanish.

E-WRITE AND THE FUTURE OF WRITING PLACEMENT IN THE VCCS

The speed and convenience of the e-Write test fulfills certain needs in a customer-service model of placement. However, the e-Write version tested here tended to pile samples into barely distinguishable masses at a few points in the rating scale—an insurmountable practical barrier to its acceptance at Reynolds for the time being. ACT is developing a version of e-Write using a 12-point overall scale that may answer that objection. If it succeeds, faculty can expect renewed pressure to adopt a single test that makes student transitions easier through a uniform placement measure.

However, writing faculty see placement through a lens that finds usefulness in the work of creating and maintaining a placement instrument. In addition to the honoring of humanistic values Drechsel (1999) identifies, conducting writing placement forces faculty to revisit vital questions: what are the basic skills of writing? What traits do we agree to recognize as demonstrating competence in these skills? Are argumentative contexts the best or only ones for eliciting the best examples of students' performance? For faculty, the work of placement may be a pearl-producing irritant; the answer to computerized testing may forever be "not yet."