Language Arts Journal of Michigan

Volume 8
Issue 1 *Historical Perspectives*

Article 6

1992

Early English Language Arts Evaluation and the Evolution of "New-Type" Tests

Ellen H. Brinkley

EARLY ENGLISH LANGUAGE ARTS EVALUATION AND THE EVOLUTION OF "NEW-TYPE" TESTS

Ellen H. Brinkley

During the last few years almost every professional education journal has devoted at least one issue to the topic of assessment. "Alternative assessment" and "authentic assessment" are currently popular expressions which underscore the fact that change is now the order of the day.

Almost always what we are seeking an "alternative" to are standardized and objective tests and the influence they have on our students and our classrooms. When we speak of more "authentic" assessment– e.g., portfolios, self-evaluation, or holistic scoring– we are rejecting what we now realize must have been the "unauthenticated" or "false" assessments of the past. More specifically, most of us resist objective, product-centered assessments designed simply to yield numerical scores that can be manipulated for a variety of administrative and sociopolitical purposes that have little to do with improving learning and teaching.

In our darker moments, we find ourselves wondering how we got ourselves into today's testing quagmire and wondering what it might take to extract ourselves from it. An historical look at early English language arts evaluation and assessment allows us to trace the evolution of our current condition and perhaps anticipate problems in the future.

Testing and evaluation were a part of the educational process in this country even in the colonial period (1607-1776). Inherent in the colonists' Protestantism was the doctrine that individuals were responsible for their own salvation and thus had to learn to read and interpret scriptures for themselves (N. Smith 11). Evaluation of reading skill– at least of its surface features– occured in the form of oral reading of the Bible or the *New English Primer* as well as by saying aloud the letters of the alphabet and syllables as listed in the primer. Clifton Johnson's *Old-Time Schools and School Books*

explains that the local minister also played an important part in evaluation. As a town officer he "examined the children in the catechism and in their knowledge of the Bible" and carried out what must have been one of the country's first evaluations of listening skills by questioning students "on the sermon of the preceding Sunday" (24).

Eventually laypersons in the community took on the task of testing student performance and of holding parents accountable for their children's learning. An educational law enacted in Massachusetts in 1642 charged "selectmen" in each town with determining "whether or not parents and masters were following their obligations," that is, determining if the children were being taught "to read and understand the principles of religion and the capital laws of the country" (Cohen 44). The stakes, at least as set by the law itself, were fairly high, for fines could be assessed parents who refused to have their children examined. If a court or magistrate agreed with the selectmen that particular parents were remiss in educating their child, the child could be apprenticed, in which case the master of the "deficient child" would be required to fulfill the provisions of the law. In 1690 Connecticut passed a similar law which made it "incumbent upon local jurymen to examine the reading ability of all the town's children" and to fine negligent parents (Cohen 81). Cohen points out , however, that in actual practice parents and towns often found ways around the penalties and that sometimes the student readers' only test was to recite a memorized catechism, a task which did not actually measure reading skill at all (81).

By the mid-1700s prospective students of Benjamin Franklin's English School ("English" in this case used to distinguish the school from those emphasizing Latin and Greek) had to meet the following entrance requirements: "It is expected that every Scholar to be admitted into this School, be at least able to pronounce and divide the syllables in Reading, and to write a legible Hand . . ." (qtd. in W. Smith 177). These may sound like modest criteria, but Franklin expressed definite ideas about more than surface features. In describing the second of six classes to be taught, he complained that the boys

> . . . often read as Parrots speak, knowing little or nothing of Meaning. And it is impossible a Reader should give the due Modulation to his Voice, and pronounce properly, unless his understanding goes before his Tongue, and makes him Master of the Sentiment. (W. Smith 179)

Writing lessons, however, focused throughout the early years primarily on penmanship and spelling, and evaluation was probably dependent on what could be demonstrated for all to see. The emphasis on good penmanship is indicated by "exhibition pieces" which were passed around for visitors to admire on the last day of the school term (Johnson 112). Such pieces were presentable, however, only after the teacher had judged them meritorious: "All their letters to pass through the Master's Hand, who is to point out the Faults, advise the Corrections, and commend what he finds right" (W. Smith 181).

During these years when many in the general population did not read or write and when in many homes the only book was the Bible, oral language was considered especially important. Although relatively few went to college, those who did found that the colleges focused great attention on rhetoric and oratory, following the "oral-based eighteenth-century model of education" (Lunsford 3). The ability to speak correctly and persuasively in public could be easily evaluated by student performance. Oratory made demands on listeners as well, though early educators seemed less concerned about evaluating listening.

As the country's attention shifted in 1776 to revolution and independence, the explicitly religious emphasis in classrooms was replaced by a nationalistic and moralistic emphasis (N. Smith 37) that affected English language arts evaluation as well. It was hoped that reading would foster loyalty in the new nation as well as "high ideals of virtue and moral behavior" (N. Smith 37). The primary reason literature was added to the curriculum, however, was so that it could serve as a subject upon which composition assignments and examinations could be based (Applebee 30). Writing also eventually took on new importance as a way to the use of written rather then oral recitation examinations because written tests were thought to be more objective:

> . . . written exams provide all students the same question in the same setting. Oral examiners necessarily had to ask different questions during their turns. Oral examiners also could phrase their questions so that some answers were more obvious than others. As a result some students received easy questions, while other students received difficult ones. (qtd. in Moore 958)

Written test responses, then, were adopted because they seemed more fair to student test-takers, though testing efficiency was surely another factor. Seen from the historical perspective, this adjustment is an early case of the English language arts curriculum itself being changed at least in part as a way to facilitate testing.

Written college admissions tests eventually became a hotly-debated topic for both high school and college English faculties. Once again the curriculum was changed because of testing as the high schools struggled to match their curricula to the college reading lists so that their students would not be at a disadvantage when they faced their entrance exams. High school teachers became exasperated, however, when faced with the need to teach so many works of literature listed by so many different colleges and when they realized how little control this left them over their own curricula. Ultimately, it was the high school teachers' complaints that led to the formation of the National Council of Teachers of English in 1911 and to the advent in 1912 of the *English Journal* (Hook 14).

Because of the college entrance examination controversy, evaluation of student performance was an important subject from the first issue of the *English Journal.* Soon, however, the pages of the *English Journal* focused less on the college entrance debate and more on description and discussion of the "new-type" objective tests and on new theories about how student evaluation could and should be handled.

Such a shift made sense to the many turn-of-the-century educators who were concurrently placing less faith in God and religion and more in what were thought to be scientific "truths." The promise of objective truth was welcomed in almost every quarter, and the "new-type" tests were quickly put into place with little regard for their profound consequences.

Perhaps the most controversial form of these "new-type" tests was the intelligence test. Intelligence tests were being developed which would eventually have far-reaching influence in the schools. The Binet scales (1905-8) and the Stanford revision (1916) were used during World War I in what today would be labeled an extremely "high-stakes" assessment situation, for the tests were used to classify recruits to determine who would serve in leadership positions and who would be sent to the front lines. The real-life tests that occurred during the fighting of World War I revealed other important results—for example, that thousands of soldiers could not read

well enough to follow printed military instructions (N. Smith 158). As educators resolved to improve students' reading skill and education in general, they soon realized the "tracking" potential for IQ tests in the schools, i.e., the grouping of "similar" students to match instruction to students' abilities (Applebee 82).

Among the early issues of the *English Journal* were articles which highlighted the need for more accurate and especially more expeditious ways to evaluate student performance. This topic was right on target for school teachers and administrators, who had seen elementary and secondary student populations jump from almost 7 million in 1870 to almost 18 million in 1910. They also had seen the number of high schools increase from 500 in 1870 to an amazing 10,000 just forty years later (Kirschenbaum, et al. 51). English teachers eventually were faced, then, with classes as large as 50 students and more, with the result that– especially for high school teachers who met a new group each hour– it was extremely difficult to know students individually or to read individual essay exams. In these circumstances, objective and standardized tests were appealing indeed.

At the same time the size of student populations seemed to create the need for objective measures, educational leaders set out to demonstrate the unreliability of individual classroom teachers' evaluative judgements. Ernest Noyes, for example, optimistically called for a "clear-cut, concrete standard of measurement which will mean the same thing to all people in all places and is not dependent upon the opinion of any individual" (534). Starch and Elliott reported a study of essays that had been graded by teachers in 142 schools (cited by Kirschenbaum, et al.). They found that one particular paper had been scored from 64 percent to 98 percent while another had been scored from 50 percent to 97 percent. Another student's paper had been given failing marks from 15 percent of the teachers while 12 percent of the teachers had given it a score of 90 percent or above. Kirschenbaum et al. explain that "with more than 30 different scores for a single paper and range of over 40 points, there is little reason to wonder why the reporting of these results caused a 'slight' stir among educators" (54-55). Given these conditions and concerns, then, it is not surprising that "scientific," i.e., standardized and objective, tests soon captured the attention of English and language arts educators at all levels.

Assessment concerns had English educators struggling to meet the needs of teachers who believed it was their weekly duty to assign and correct

students' product-centered compositions. How to test composition efficiently seemed to pose the problem commanding the greatest attention among English educators and led to the development of a number of composition "scales" intended to provide an objective set of standards by which to evaluate writing. One of the first and one of the most popular was the Hillegas scale. A 1912 *English Journal* article explained how the scale had been developed: A large number of student compositions had been sent to several hundred judges, who were asked to arrange the papers in order of merit. From these rankings, a scale of ten samples "ranging in value by equal steps from 0 to 937 units" was derived (Noyes 535). (Actually the zero point was established on the basis of an "artificial sample produced by an adult who tried to write very poor English" [Noyes 535], an understandable cause for later criticism of this particular scale.) The ten sample papers and their percentage scores were copied and distributed to serve as what today would be called "range finders" by teachers who could compare their own students' writing to the samples.

It is interesting to notice that the test designers projected several other benefits that might result from using such measures. However, rather than focusing on what the tests could tell teachers that might ultimately improve instruction, test designers pointed out that supervisors could use the samples to "compare classes of the same grade in different schools, in different cities, or under different teachers" (Noyes 536). These suggestions, then, emphasized the external administrative uses that could be made of test scores and at least implied the possibility of linking teacher evaluation to student performance on the basis of what were thought to be objective measures.

Not everyone bought the notion of writing scales. C. H. Ward in a 1917 *English Journal* article, "The Scale Illusion," attacked the practice of ranking themes (221), arguing that "[a]ny measure of literary value is impressionistic; any measure of literary value and mechanical value at the same time is a phantom" (223-4). He further insisted, "A system that shows [the student] only his height above an absolute zero can no more produce a harvest than a thermometer can bring forth figs" (230). At first glance, Ward seems to offer a refreshing emphasis on students' growth and learning. However, what Ward offered instead of the Hillegas scale would not be well received by modern educators, for it was a system based on the principle of subtraction for errors from a perfect score—as if correctness was all that mattered.

This trend toward objective testing was taking place in the other language arts as well. Starch's 1916 book included a reading test he designed that may have served as an early model for later standardized reading tests. It was intended to measure the "chief elements" of reading, perceived by Starch as comprehension, speed, and correctness of pronunciation (20). He offered several reading passages at increasingly difficult reading levels, which students were asked to read silently for thirty seconds. After the reading, they were asked to mark the spot where they stopped reading and to write down as much as they could remember from their reading. Interestingly enough, the written retelling was scored by crossing out the words which reproduced the text and by counting those remaining– seeing what percentage of words should be discarded as not related to the text (31). Somewhat similarly, Gray's oral reading tests asked the child to read aloud while the tester recorded the errors made, the idea being that the better students could read faster with fewer errors (Stone 263). Both tests, then, focused more on surface reproduction of text than on meaning.

Even more problematic were test items phrased as rather naive yes-no questions, such as "Are men larger than boys?" It is easy to imagine young test-takers thinking to themselves, "Yes, men are usually larger than boys, but . . . ." Modern test designers might similarly object to the Burgess Rate Test, which sometimes asked students to draw a response:

I. This naughty dog likes to steal bones. When he steals one he hides it where no other dog can find it. He has just stolen two bones, and you must take your pencil and make two short, straight lines, to show where they are lying on the ground near the dog. Draw them as quickly as you can, and then go on.

Figure 1 Sample Unit of the Burgess Silent-Reading Test (Stone 238)

Again, we can imagine test-takers' anxiety as they tried to decide the "right" place to draw the bones.

Other tests were discussed and developed during this era of test fascination and explosion– some of which look very unrealistic to readers today. For example, the Ayers handwriting scale, as described by Starch, was intended to evaluate students' penmanship. Incredibly, the test was constructed by taking samples of 1,578 children's handwriting, separating the individual words, and then measuring the speed with which readers could read these words. Eventually eight degrees of legibility were determined and presented to be used as guides, with three samples of each– slant, medium, and vertical. The following figure shows a portion of the scale:

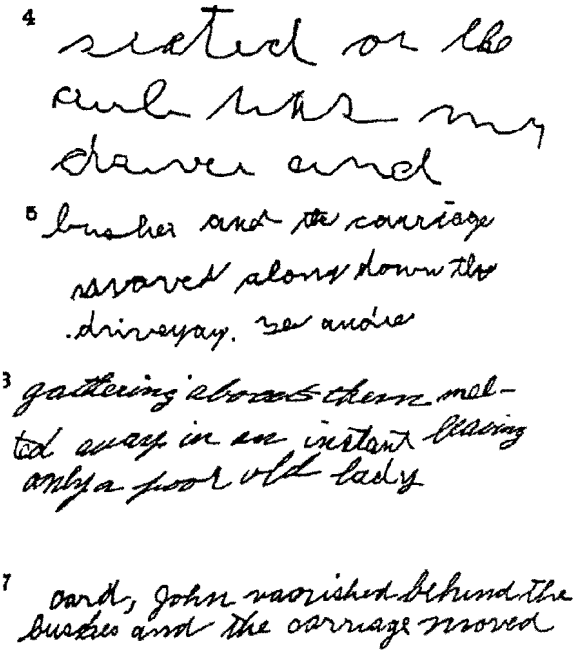

Figure 2 Ayres Handwriting Scale (Starch 62-63)

It is difficult to imagine the motivation for such excessive and obsessive procedures, but such measures seem to reflect the enormous faith that

educators and the general public had in ultimately objectifying every part of their existence.

Given this trend, it is no surprise that literature posed unique evaluation problems that were discussed on occasion but were difficult to address effectively. Efforts were made, however, to create standardized tests of appreciation of literature. Leonard has described the process used in such a test that was intended to measure the literary appreciation of both teachers and students. The test presented a number of poems "ranging in quality from Mother Goose to Bridges or Masefield." Each poem was accompanied by three "spoiled" versions, and test takers were asked to determine which in each case was the "best" (59), thus supposedly demonstrating the ability to discern and appreciate real literature.

Throughout this period some English educators called for standardized tests to be developed in even more areas. For example, Klapper sought a scale for oral composition, since he believed it was much more important than written composition, which he termed "incidental" in the life of the average person (221). By 1925 early issues of *The Elementary English Review* duly reported projects in Detroit and Chicago to develop oral composition standards (Hosic; Beverly).

Some educators early in this century, however, did focus on evaluation specifically for the purpose of improving classroom learning and reminded readers of professional publications that"desirable language habits" were best observed in everyday oral and written expression. Without using the expression "reflective practice," they encouraged teachers to think of their students' demonstrated skill to revise teaching practices (Savitz, Bates, and Starry 2).

Other authors of this period made evaluation recommendations that eventually became established classroom practice. For example, a 1918 *English Journal* article echoed what Franklin had recommended two centuries earlier by arguing against "old-time memory tests" that asked students to parrot back information provided by texts and teachers. The author offered open-book "thought examinations" instead (Wiley 327). For example, tenth grade students who had read *A Tale of Two Cities* and *The Merchant of Venice* might be asked, "Do you think Shylock was more or less vindictive than Madame DeFarge? Explain" (329). Interestingly, Wiley points out that this form of testing had a positive effect on the teacher's subsequent teaching:

> She found . . . that clear thinking by a pupil on examination required inspirational teaching on the part of the teacher day by day. (327)

Although some questioned the validity and reliability of composition scores and standardized reading tests, most educators during the early- to mid-1900s seemed to side with Daniel Starch, who argued that "[a]ny quality or ability of human nature that is detectable is also measurable" (2). Starch recommended that the test results be used to develop "a definite standard of attainment to be reached at the end of each grade" (31-32). Such a standard, he insisted, would make it possible for a

> qualified person to go into a schoolroom and measure the attainment in any or all subjects and determine on the basis of these measurements whether the pupils are up to the standard, whether they are deficient, how much, and in what specific respects. (32)

Significantly, Starch said little, however, about where the standards might be set, who might set them, or why.

Caught up in the testing spirit, by the mid-1920s most English language arts educators seemed convinced that numerical scores from the "new-type" tests were the best means by which to set standards for English language arts student performance. Indeed, objective measures of students' ability and achievement were hailed as "the most significant movement in education during the 20th century" (Thomas 438). Finally, NCTE, which had aired so many of the pros and cons in the *English Journal*, spoke out on the testing issue. Today's readers might be surprised to learn that the June 1923 issue included a report from the NCTE Committee on Examinations, whose first sentence made it clear where the professional organization stood at that time: "The Committee on Examination desires to stimulate an interest in a more widespread use of standard tests in English" (Certain 365). Apparently, NCTE, which has in recent years rejected widespread use of standardized tests, took the 1923 stand because of what they perceived at that time as positive value in allowing school districts to compare their test scores with those from other districts (Certain 365).

Although we may be tempted to chuckle at the naivete of both NCTE and early test-designers, it is easy to understand the promise that standardized scales and tests held—to turn students' reading and writing into numerical scores not tainted by human attitudes and impressions. The same impersonal efficiency that seemed to work on the factory assembly lines promised both higher productivity and quality control in English language arts evaluation as well. Few seemed, in print at least, to question the effects of sorting students on the basis of numerical labels. Few seemed to question whether the tests themselves could accurately and adequately evaluate the complexities of English language arts skills and processes.

We notice with the wisdom of hindsight, then, just how much faith English language arts educators placed in the tests. In spite of all that's published today about assessment, there's no guarantee that we won't continue to repeat the mistakes of the past. We know that today's English language arts assessment theory is student-centered and process-focused and intended to describe students' strengths as well as weaknesses so that classroom teaching and learning can be improved. Although it's difficult to think through all of the possible ramifications of the assessment measures being recommended today, history shows us that we would do well to be as informed as possible and to seek multiple measures when the stakes for literacy learning are so high.

Works Cited

Applebee, Arthur N. *Tradition and Reform in the Teaching of English.* Urbana: NCTE, 1974.

Beverly, Clara. "Standards in Oral Composition: Grade One." *Elementary English Review* 2 (1925): 360-61.

Certain, C. C. "Are Your Pupils Up to Standard in Composition?" *English Journal* 12 (1923): 365-77.

Cohen, Sheldon S. *A History of Colonial Education, 1607-1776.* NY: John Wiley, 1974.

Hook, J. N. *A Long Way Together.* Urbana: NCTE, 1979.

Hosic, James F. "The Chicago Standards in Oral Composition." *Elementary English Review* 2 (1925): 170-71.

Johnson, Clifton. *Old-Time Schools and School-Books.* 1904. Intro. Carl Withers. NY: Dover, 1963.

Kirschenbaum, Howard, Rodney Napier, and Sidney B. Simon. *Wad-ja-Get?* NY: Hart, 1971.

Klapper, Paul. *Teaching English in Elementary and Junior High Schools.* NY: D. Appleton-Century, 1915.

Leonard, Sterling Andrus. *Essential Principles of Teaching Reading and Literature.* Philadelphia: J. B. Lippincott, 1922.

Lunsford, Andrea A. "The Past—and Future—of Writing Assessment." *Writing Assessment.* Eds. Karen Greenberg, et al. NY: Longman, 1986. 1-12.

Moore, David W. "A Case for Naturalistic Assessment of Reading Comprehension." *Language Arts* 60 (1983): 957-69.

Noyes, Ernest. "Progress in Standardizing the Measurement of Composition." *English Journal* 1 (1912): 532-36.

Savitz, Jerohn J., Myrtle Garrison Bates, and D. Ralph Starry. *Composition Standards.* NY: Hinds, Hayden & Eldredge, 1923.

Smith, Nila Banton. *American Reading Instruction.* 1934. Newark, DE: IRA, 1965.

Smith, Wilson, ed. *Theories of Education in Early America 1655-1819.* Indianapolis: Bobbs-Merrill, 1973.

Starch, Daniel. *Educational Measurements.* NY: Macmillan, 1916.

Stone, Clarence R. *Silent and Oral Reading.* Boston: Houghton Mifflin, 1926.

Thomas, Charles Swain, et al. *Examining and Examination in English.* Cambridge: Harvard U Press, 1931.

Ward, C. H. "The Scale Illusion." *English Journal* 6 (1917): 221-30.

Wiley, Mary Callum. "The English Examination." *English Journal* 7 (1918): 327-30.

Ellen Brinkley is an Associate Professor of English at Western Michigan University and President-Elect of the Michigan Council of Teachers of English.