Writing to the Rubric

Lingering Effects of Traditional Standardized Testing on Direct Writing Assessment

BY LINDA MABRY

Performance assessment offers a way out of the factory and into the classroom, the author points out. But to get there, we need to reconsider our mania for making test-based comparisons that demand standardization.

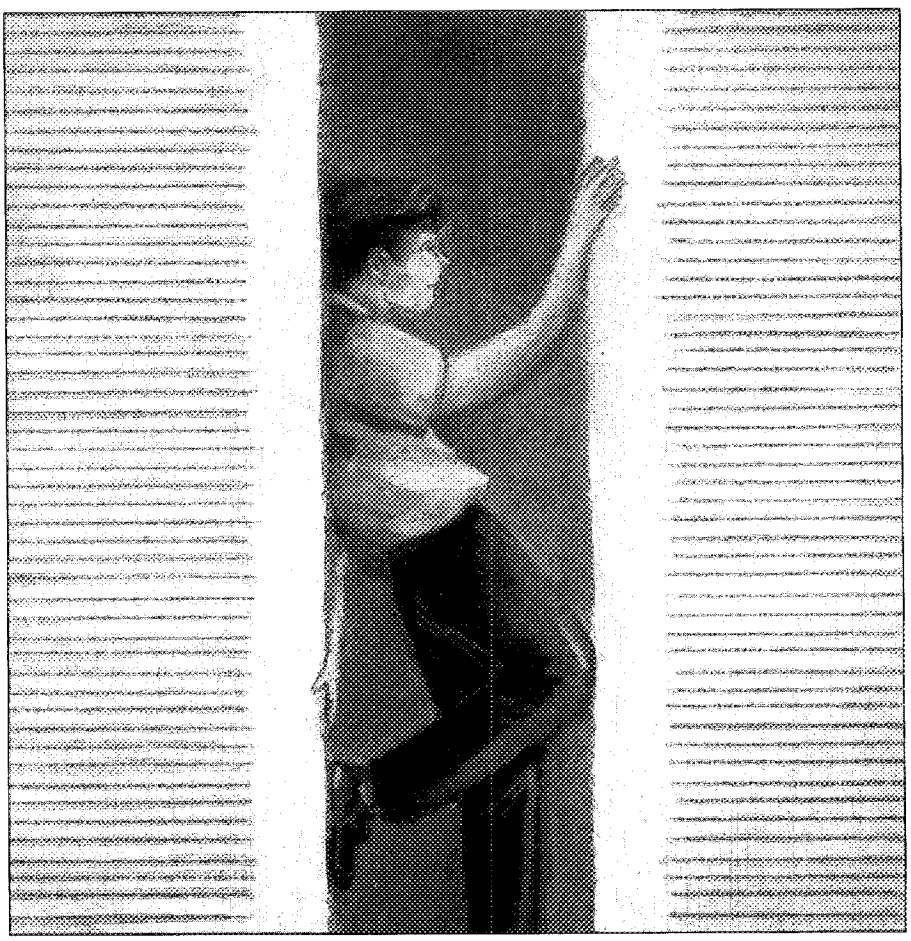


Illustration by Karen Stolper

AS SHE placed test materials, dictionaries, and thesauruses strategically on empty desks in preparation for administering Pennsylvania's state-mandated direct writing assessment, a middle school teacher awaiting her students said:

> See this box of laminated rubrics? That's the rubric the state uses to score this test. I also score the writings of my students in this class with it. Sometimes I have students use it to score each other's papers. They have to justify the scores they give, so they are very familiar with this rubric. They know what they have to do today when they take the test.
>
> But the test is not really a fair representation of how well they can write. They could do better if they had a choice of topics and could do the sorts of things we usually do. For instance, with these papers over here, they picked a topic one day in advance and could think about it, look at some reference materials, take some time planning and drafting, and come up with a polished product. I was very pleased with these papers.
>
> I'm frustrated with the test. We've done all of this preparation, and I've organized to the max so they can concentrate on their writing during the test time, but my students will still score at only about the state average. We're a rural district, and we don't have all the curricular options you find some places. I have to teach literature and grammar as well as writing, and I throw in spelling — they're not always going to have a spell-checker handy. But my students will be compared to some suburban students who are in a writing-only curriculum. When the scores are printed in

LINDA MABRY is an assistant professor of educational psychology, Indiana University, Bloomington.

Copyright © 1999. All rights reserved.

the newspaper, people will think we're not doing a good job of teaching here. It just doesn't seem fair.[1]

The direct assessment of student achievement in writing — judging writing skill on the basis of actual student writing instead of multiple-choice test items — was supposed to fit better with daily classroom activities and objectives and was supposed to provide a better basis for making valid inferences about student learning. So what has gone wrong when a state-mandated performance assessment is threatening curriculum, creating anxiety, and raising equity concerns?

As part of the larger argument that traditional psychometric concepts and practices undermine new ideas and techniques in assessment, I will argue that rubrics have the power to undermine assessment. Scoring rubrics are pivotal in operationalizing large-scale and standards-based performance assessments in writing. Rubrics promote reliability in performance assessments by standardizing scoring, but they also standardize writing. The standardization of a skill that is fundamentally self-expressive and individualistic obstructs its assessment. And rubrics standardize the teaching of writing, which jeopardizes the learning and understanding of writing.

Operationalizing Standards In Rubrics

Of the performance assessments in state testing programs, direct writing assessment is the most prevalent and longest-running.[2] Currently, 38 states assess students' writing skills through direct writing assessments, and all 38 use a rubric to score students' performances.[3]

Rubrics. Frequently referred to as scoring guides, assessment rubrics are rules by which the quality of answers is determined. Each rubric is a "description of student performance that clearly articulates the requirements for each of the score points."[4] Rubrics are translations of visions of desirable performance into specifications of exactly what is desirable. The use of rubrics to score performance assessments predates the current standards movement, but rubrics are crucial in current implementations of academic standards.

Rubrics are artifacts of the psychometric tenet that good assessment begins with careful thinking about what a test-taker should know and how that knowledge should be appraised.[5] In standardized, norm-referenced, multiple-choice testing, agreements about what constitutes achievement are operationalized in a table of specifications and then in a test.[6] Similarly, in large-scale performance assessment, agreements about what constitutes achievement are operationalized in rubrics and in performance items of various types. The strategy is *predictive* in that the performance of test-takers is anticipated, and it is *preordinate* in that what will count as satisfactory performance is determined before the test is administered. This is an orderly, linear conception of test development. But current practice of standards-based assessment presents a contrastingly complicated scene with a tangle of ideas about standards and rubrics.

Standards. A decade ago, the widely admired standards of the National Council of Teachers of Mathematics (NCTM) declared consensual notions about what and how mathematics should be taught and about which student understandings and skills should be promoted in educational settings.[7] The success, utility, and impact of the NCTM standards motivated educators in other subject areas to develop their own content standards, often in areas, such as history/social studies and English/language arts, where there was less consensus.[8] At the federal level, when receipt of Title I funding was made contingent upon reporting of student achievement by performance standards, 35 states were granted waivers in order to develop state standards by September 1997.[9] Currently, every state save Iowa has adopted or is currently developing standards.[10] Like rubrics, standards resonate with traditional measurement because they involve advance thinking about what should be assessed and how.

At the conceptual level, a distinction is drawn between content standards, performance standards, and delivery or opportunity-to-learn standards. Robert Linn[11] has stated the general understanding of these terms: Content standards articulate the concepts and skills students should be taught and should learn; they tell what to assess. Performance standards describe more explicitly and more concretely what students should demonstrate in order to show proficiency. Delivery standards, arguably the most important in terms of equity,[12] are the most problematic to implement;[13] they define the quality of educational opportunities and resources that must be provided to students. In the abstract, the differences between these types of standards are clear.

Crippling ambiguities. At the 1997 meeting of the American Educational Research Association (AERA), several nationally prominent measurement experts and others well informed about state-level assessment were asked: Is there a difference *in practice* between content and performance standards? Or does specifying performance levels for the purpose of scoring in effect transform content standards into performance standards? At the conceptual level, the definitions are clear. But in an ongoing study of state-mandated performance assessment, when personnel in state education agencies were asked whether their states had content or performance standards, respondents frequently answered that their states had content standards but sent the researchers copies of performance standards or rubrics.[14] Some state standards appeared to be amalgams, statements of *what* it was hoped students might learn (i.e., content standards) interlaced with statements of *how well* students were expected to perform (i.e., performance standards). The experts uniformly confirmed that the distinctions had indeed become blurred in practice.

An insignificant matter, perhaps. If we have good assessment and are able to communicate with one another about it, what we call our standards might be immaterial. But close examination reveals that, among important players in high-stakes contexts, common measurement terms are not similarly used and sometimes not understood. Communication is obstructed to such an extent that answers to the most basic questions on carefully constructed surveys fail to elicit accurate information even on a matter as straightforward as whether or not a state has implemented performance assessment. The shallowness of understanding about assessment in some state education agencies suggests an alarming dependence on test development companies and the need for response from the measurement community.[15]

Confusions in direct writing assessment. Using the term *standard* at two levels invites confusion. Writing rubrics typically specify both *what* students should know and *how well* they should perform. Criteria tell the what. The how well is variously called standards, performance levels, or performance standards. Finer grained than the more superordinate content stan-

Copyright © 1999. All rights reserved.

dards or performance standards, the standards in rubrics are specific descriptions of performance levels, stipulations of details expected in the performances of test-takers, or categories into which test-takers' performances are grouped. For example, Vermont's "Analytic Assessment Guide" for writing portfolios lists its criteria as 1) purpose, 2) organization, 3) details, 4) voice/tone, and 5) usage, mechanics, grammar. It lists its standards as 1) extensively, 2) frequently, 3) sometimes, and 4) rarely.[16] Kentucky's "Holistic Scoring Guide" for writing assessment lists its criteria as 1) purpose/audience, 2) idea development/support, 3) organization, 4) sentences, 5) language, and 6) correctness. It lists its standards as 1) distinguished, 2) proficient, 3) apprentice, and 4) novice.[17]

Often, a rubric is presented as a matrix in which the criteria and the standards form horizontal and vertical axes. For each cell in the matrix, an anchoring description may be provided and a student paper offered as a benchmark to help scorers match performance levels for each criterion with test-takers' writings. Some rubrics are presented as check lists or performance levels that have both criteria and standards embedded.

The use of the term *holistic* to describe some rubrics also suggests confusion. Holistic evaluation prioritizes the total effect of a piece of writing, the irreducibility of the whole, in contrast to analytic evaluation, which highlights separate components and might perhaps be followed by an evaluative synthesis of these constituent elements. In specifying the criteria by which student performance will be judged, writing rubrics focus attention on those components more than on the overall effect of the writing.

Producing a single score appears to be enough to satisfy rubric developers that a score is holistic. But when the awarding of an overall score is the result of aggregating subscores on components, the single final score is a product of analytic scoring. Although it is possible to derive a holistic score from a list of discrete criteria, strategies that focus on criteria are fundamentally analytic in character because they focus on components. Because rubrics to assess writing prescribe the criteria by which papers are to be judged, claims of their holism rarely survive analysis.

It is not semantic hairsplitting to differentiate between holistic scoring and analytic scoring. One implication of the failure to distinguish between the two is that traditional large-scale achievement testing has been so analytic for so long that even a basic understanding of a holistic approach appears to be missing. That holistic scoring might be an improvement over analytic scoring is apparent from two observations regarding evaluative check lists: first, the criteria on a check list must predict every important feature of performance (although it is doubtful that the hundreds of thousands of individual performances elicited by large-scale testing can be adequately predicted); second, failure to weight items appropriately on the check list yields invalid judgments.[18] When test scores carry such high stakes as school takeovers, the loss of jobs, and the denial of diplomas and opportunities for higher education, invalidity is intolerable.

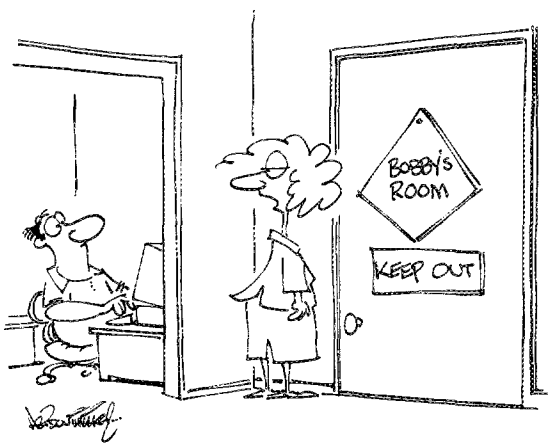
Validity and Reliability With Rubrics

Rubrics offer both explication and complication. Rubrics tend to improve *interrater reliability*, the likelihood that different raters will award similar scores. But the consistency is not achieved because rubrics provide a vehicle for expressing naturally occurring agreement.[19] Rather, rubrics limit the scope of variability of scores. That is, rubrics improve interrater reliability partly by directing all scorers to judge student writing according to the same few criteria, a sameness that encourages agreement in scores. And writing rubrics typically incorporate attenuated measurement scales of three to six performance levels, which offer a scorer few choices. The fewer the choices, the fewer the possibilities for disagreement among scorers, and the fewer but more serious the measurement errors.[20] Agreement is further enhanced because scorers are trained to use rubrics uniformly and are monitored as they do so, their continued employment contingent on a record of awarding consistent scores.

Although rubrics promote reliability, they may simultaneously undermine validity, the more important determinant of the quality of an assessment. Writing rubrics can fail to predict the actual features of a student's writing, thereby creating a mismatch between scoring criteria and actual performance. In cases in which the overall effect of a student performance is achieved by means not anticipated in the scoring criteria, criterial analysis of the quality of writing will deflect a scorer's attention away from the actual writing, and the score will not support valid inferences about the student's achievement.

In traditional measurement thinking, reliability is supposed to support validity. Measurement textbooks uniformly proclaim that reliability is necessary but not sufficient for validity. But consistency among scorers can reflect collective tunnel vision rather than informed consensus about the quality of student writing. Forced attention to features anticipated by a rubric rather than to actual features of student work can yield inappropriate or inaccurate scores and lead to invalid inferences regarding student achievement.

Domain appropriateness. Rubrics present another potential threat to validity in the assessment of writing. Prespecification of scoring criteria is consistent with traditional principles of test development



"Honey, e-mail Bobby and tell him dinner's ready."

Copyright © 1999. All rights reserved.

— the long-standing practices of articulation of constructs, purposes, topics, formats, and difficulty levels before test construction — and with the traditional practice of standardizing across test-takers and contexts. But complying with prespecified criteria and common standards works against creative self-expression, which is the essence of skillful writing. Because they restrict the flexibility scorers need in order to identify and commend unique strengths and skills, rubrics used for scoring large-scale, standardized performance assessments in writing not only undermine validity but are fundamentally domain-inappropriate, not sufficiently relevant to and representative of the domain of writing.[21]

Choice of criteria. Writing rubrics typically feature four to six criteria that give priority to the mechanical, format, and organizational aspects of writing rather than to the more substantial aspects such as content, logic, compelling presentation, vivid description, figurative use of language, depth of character, or significance of theme. It may be that low-level criteria dominate writing rubrics because agreement among scorers is more easily achieved with regard to such matters as spelling and organization than to less tangible considerations, such as whether the writing is persuasive or whether the metaphors deepen a reader's understanding. At least one empirical study has shown that attention to form-dominated criteria on Texas writing rubrics blocked attention to the substance of writing when the state's benchmark for a perfect paper failed to address the specific topic stated in the writing prompt.[22]

Imposition of prompts. Rubrics are sometimes used to assess writings collected in portfolios. In large-scale programs, however, direct writing assessment is more often "on demand." That is, the assessment provides students with a "prompt," a topic or question to which students are to respond within an immediate testing period. The Religious Right has objected to prompts eliciting a degree of student self-revelation. Consequently, procedures for developing prompts typically take great care to include public input and to avoid any prompts that might be construed to be invasions of student privacy.[23]

Yet by avoiding topics involving students' values, perspectives, or accounts of personal activities, some prompts have become so spiritless and bland that it is difficult to imagine that students might be motivated to do their best writing in response.[24] The imposition of both an external prompt and external criteria can introduce a deadly distance between the writer and the subject, denying a young writer authority over the most basic aspects of the writing task — what to write about and how to make the writing good. This ruinous combination raises a serious question of validity: a student's achievement cannot be assessed if the task elicits a lackluster, indifferent response that does not reflect the student's achievement or notions of quality.

Importance of reliability. In our increasingly high-stakes testing environments, reliability is necessary to promote confidence that test scores properly facilitate the distribution of privileges and resources. Reliability is also necessary to preempt litigation regarding the property rights associated with score-based decisions such as the denial of high school diplomas.[25] Not coincidentally, there is growing presumption that good performance assessment requires a rubric. So scoring rubrics have become pivotal in operationalizing large-scale and standards-based performance assessments, typical in scoring not only writing but all types of constructed-response items.

Today's complex environment for performance assessment is freighted with irony. Rubrics discourage attention to the very student writings that direct writing assessments elicit. And as agents of standardization, writing rubrics make direct writing assessments more like the multiple-choice tests they were meant to improve upon, thus rendering "alternative" assessment something of a misnomer.[26]

The Monitor Becomes the Goal

In 1921 Edward Thorndike, who has been called the father of educational measurement, responded to criticism of newly developed standardized tests:

> It will be said that learning should be for learning's sake, that too much attention is given already in this country to marks, prizes, degrees and the like, that students work too much for marks rather than for real achievement. . . . Students will work for marks and degrees if we have them. We can have none, or we can have such as are worth working for.[27]

Today, with educational accountability grounded in high-stakes assessment, the prescience of Thorndike's early critics is evident: test scores have come to dominate education. Not every student (and not every educator) thinks the scores are "worth working for," but, threatened with high stakes, few dare to act as if learning were more important than test scores.

In traditional measurement, testing is routinely described as supporting education, monitoring student progress, informing teachers, and helping them plan curriculum and instruction. The implication, at least, is that a student's test results should facilitate recognition of his or her level of achievement and should help target instruction. But standardized achievement testing has proven not very diagnostic, not very sensitive to individual attainment, not very helpful in prescribing appropriate remediation, and not very supportive of teaching or learning. Instead, contemporary critics have condemned the tests for bias, distortion of curriculum, misalignment with current learning theory and best practice in pedagogy, misallocation of educational resources, deprofessionalization of teachers, demoralization of students, and entrenchment of social and political inequities.[28] To the extent that new assessments fail to escape the force field of old measurement ideas and practices, they are subject to the same charges. High stakes concentrate attention on test scores and lock in these negative consequences. Production of test scores overwhelms education as the mission of the school.

Effects of Rubrics on Teaching

Similarly, as the crucial mechanism for determining scores in direct writing assessment, rubrics overwhelm the writing curriculum. The powerlessness of educators to resist the implacable influence of state testing in general and of writing rubrics in particular was apparent in observations of local administration of state-mandated performance assessments in Pennsylvania, Michigan, and Indiana in 1998.[29] Teachers know, of course, that they will be excoriated if they succumb to pressure to "teach to the test" — a familiar accusation that many teachers interviewed in these three states repeated and denied. But evidence to the contrary was unmistakable. In some classrooms, state writing rubrics were accorded unprecedented priority, as indicated in the extended quotation with which I opened this article.

Copyright © 1999. All rights reserved.

In other cases, the influence of tests and rubrics was more subtle. An urban elementary school principal in Indiana explicitly denied that state performance assessments constrained local teaching but nevertheless admitted, "Consciously or unconsciously, our teachers probably teach in a way that will help students perform well on the test." She elaborated:

> This year, for the first time, we received information for teachers — one book for grades 1-3 and another for grades 4-6 — about teaching methods to help raise scores, and a book to help with remediation for math and reading. That was a first. It was helpful. Nobody said we have to use them, but they could be valuable resources for the teachers. . . .
>
> [Our district assessment coordinator] says over and over not to use test-specific materials, that it's not ethical to do so. . . . We used to use [test-specific student preparation materials] four or five years ago, like everyone else in the state does. But there was publicity, and we were told we couldn't use [them]. We got rid of all of those materials; we got them totally out of the building. It was upsetting at first when we learned that other school corporations were [still] using those materials, but now we accept it. Nothing we do will ever be good enough for the media or for the mayor.
>
> Teachers are under a lot of stress because of [state law that provides that] teachers in our district are evaluated on the basis of test scores. Any time you put that much importance on a test, people are going to feel — whether it's reality or not — that their jobs are on the line, and they may give extra help. That's what I think happened in this case [of a student whose test scores improved from the 5th percentile in the fall to the 56th percentile in the spring], although we go through training about all of the ethical testing practices.[30]

In some cases, teaching to the test and writing to the rubric were described in the more socially acceptable language of curricular alignment. A rural high school English teacher in Michigan said that he had thought about his students' scores from previous years and had subsequently changed his teaching. Then he suffered feelings of internal conflict:

> The test is valid, so okay. But in the back of my mind, I was worrying, "Am I teaching to the test?" And I hate that. To me, if you're teaching for a test, you're not teaching. So, that's a nagging feeling.
>
> The writing test is valid; it's a great test. [But] there's a problem with grading the test. Even kids who have gotten A's in my composition classes — and I'm a very tough teacher, a very tough grader — got awful scores. The number of students who were classified as nonproficient on the test was horrendous. Many of these kids had straight A's. I don't know what the criteria are. Even our salutatorian was scored as nonproficient. I don't get it. So I'm not sure what they're looking for.[31]

When asked whether explicit efforts were made to align the curriculum with the test, this teacher said otherwise, but then he indicated that curriculum decisions at the school were, in fact, informed by test requirements and results. With better information, he said he would do more to match his teaching to the state's direct writing assessment, despite the "superficiality" of rubric criteria, which, under other circumstances, "were not things I'd pick out for teaching." He continued, "Sometimes I get the impression the rubrics were written by people who haven't been in a high school English classroom since Moses was a baby. They don't know what quality is in a student paper." He also complained that "the scale doesn't help us understand what we could do better." But the results "are printed in the newspapers, and people look at them and say we're not teaching [well]. . . . That's not true."

This was only one of many instances in which it appeared that rubric-driven writing curricula were overriding teacher-driven writing curricula that might well have been superior. It is also an example of deprofessionalization, of teachers effectively being prevented from making decisions about curriculum and pedagogy for their students and being punished for trying to do what state testing demanded.[32]

Professional development. The impact of testing on teaching was also apparent in the widespread redirection of professional development. The benign label of professional development masked the dubious legitimacy of providing training that, instead of being designed to increase teachers' knowledge of the principles and practices of measurement, was intended to help them improve students' scores on state assessments. A district administrator in rural Michigan, for example, proudly described professional development sessions in which teachers in curriculum teams at every grade level throughout the district were trained in student test-taking skills and methods for combating stress, were assigned to review sample test items and the results of previous tests, and were required to develop instructional units that corresponded to the state test. External consultants had been hired for "the most serious professional development," which was targeted to writing.[33]

A middle school teacher in rural Indiana recounted just such a session in her school:

> The principal got all the teachers who administered the ISTEP [Indiana Statewide Testing for Educational Progress] together and had them participate in their own testing session. The teachers were given test questions that were similar to the test questions found on the third-, sixth-, eighth-, and 10th-grade tests. He did this so that we could be acquainted with the types of questions the students were asked to respond to and so that we could experience some of the same feelings the students do when they take the tests.[34]

Teachers are thus trained to comply with test requirements and to make student writing conform to assessment rubrics. This is not training that encourages teachers to consider critically whether student achievement is helped or hurt by the tests or to judge whether the impact of state assessments on curriculum should be welcomed or opposed. Training teachers to conform to the tests is a mechanism for co-opting them, for undermining their capacity to resist the tests and the rubrics, and especially for raising scores, which will become public.

Public reporting. When scores are aggregated, ranked, and reported in the media, all testing becomes high-stakes testing. When high stakes are associated with test scores, educators act to protect their careers, to insulate themselves from public humiliation and their schools from funding cuts, and to safeguard their students' educational opportunities and morale. When states and districts provide training and test-preparation materials that imply that assessment results are of paramount importance, educators are encouraged to believe that aligning curricula and pedagogy to the tests is expected and desirable. When writing rubrics are provided in advance or become familiar through experience, teachers teach students to write to the rubrics.

Copyright © 1999. All rights reserved.

In this way, instruction is corrupted by assessment.

Effects of Rubrics on Learning

Often, it is in the name of equity that students are given copies of the rubrics by which their writing will be scored. Some rubrics are even printed in test booklets. On the face of it, it does seem reasonable and fair to provide students with information about the expectations that they will be asked to meet. It is less obvious that doing so permits rubrics to set the boundaries of creative expression. Rubrics are designed to function as scoring guidelines, but they also serve as arbiters of quality and agents of control. Moreover, the control is not limited to assessment episodes but influences curriculum choices, restricts pedagogical repertoires, and restrains student expression and understanding.

Research in Illinois has documented that direct writing assessment scored with a rubric yielded formulaic writing by students and simplification and homogenization of the writing process.[35] The Illinois writing rubric indicated that a student should be credited for providing several points of support for his or her thesis. But examination of benchmark essays revealed that assessors were ignoring whether the points offered by a student actually supported the thesis and whether the student's paper exhibited coherence overall. Compliance with the rubric tended to yield higher scores but produced "vacuous" writing. Performance was rewarded on the stated criteria only, and those criteria were insufficient to ensure good writing.

In declaring performance standards, rubrics both compel and constrain student performance. It is unfortunate that the rubrics in current use demand compliance to dismembered definitions of writing. Perhaps rubrics could be devised that have the comprehensiveness and flexibility to accommodate different genres, voices, and styles of writing. But perhaps writing is too personal and varied an enterprise to be amenable to scoring by rubric.

The Shackles of Psychometric Habit

The move toward direct writing assessment is a move in the direction of authenticity and validity. Clearly, standardized multiple-choice testing presents a validity problem for assessing writing. Multiple-choice items about different aspects of writing incorrectly imply that good writing is the sum of such components as spelling, vocabulary, grammar, word choice, and sentence structure and that the ability to answer multiple-choice items on these topics is a measure of the ability to write well. But standardizing scoring with a writing rubric also presents a validity problem. Rubrics incorrectly imply that good writing is the sum of the criteria on the rubric, that the criteria on the rubric are sufficient for good writing, and that writing that does not conform to the criteria on the rubric is not good.

Standardization is essential to large-scale testing. Standardization of direct writing assessments promotes reliability in scoring and facilitates the comparison of students, schools, and districts. Thus standardization caters to the insatiable public appetite for rankings. And there's the rub. Rubrics standardize scoring, and so they standardize writing. But standardized writing, by definition, is not good writing because good writing features individual expression, which is not standardized. The standardization of any skill that is fundamentally individual obstructs its assessment. And this presents a validity problem because the assessment fails to produce scores that support valid inferences about students' writing achievement. More specifically, it is a problem of construct validity — testing not the construct of *writing achievement* but the construct of *compliance to the rubric.*

If writing were scored in a truly holistic manner, using trained professional judgment, there would be less prespecification of criteria and more flexibility in scoring. Students would be evaluated more on what they actually write than on how well their writing matched the scoring criteria. But there would also be more variability in the scores. Such a situation would be authentic — in the world outside schools, critics of writing do not always agree — but psychometrically intolerable. When push comes to shove in scoring direct writing assessments, reliability takes precedence over validity.

Because evidence of validity is always incomplete[36] and because of the traditional assumption that reliability is a prerequisite for validity, test developers have long concentrated more on providing evidence of the reliability of tests than on providing evidence of their validity. Developers of large-scale performance assessment programs have likewise worked hard to produce evidence of reliability. The development of rubrics and the training and monitoring of scorers for the consistent use of the rubrics is critical to the plan. But despite concerted efforts, the record indicates that reliability in performance assessment is fiendishly difficult to achieve.[37] Caught in old psychometric habits and assumptions, developers and evaluators of performance assessments react to the difficulty by pressing still harder for reliability.

Recognition and reconsideration are needed instead. The public needs to reconsider its demand for high stakes and merciless comparisons. The measurement community needs to reconsider the importance of reliability where it is attained at the expense of validity. General recognition is needed of the destructiveness of standardization to writing and to the teaching of writing.

Education has long tried to break the bonds of the industrial metaphor. But education is heavily shackled by punitive, test-driven school reform. Other means of accountability have been offered,[38] but testing increasingly drives educational accountability and reform.[39] Now found in direct writing assessment and other types of performance assessment as well as in multiple-choice testing, standardization is partly a response to the large scale of these assessments and partly a matter of habit. Performance assessment offers a way out of the factory and into the classroom. But to get there, we need to reconsider our mania for making test-based comparisons that demand standardization. Such standardization now unfailingly involves the use of rubrics in assessing writing or of constructed responses in other subjects. Thoughtful reconsideration of these issues was apparent in the rationale offered by Nebraska when it discontinued state testing in 1996-97:

> We understand what is occurring nationally and that in many ways we may appear to be way behind other states in the use of state-centered assessment and reporting systems. . . . However, we are also absent all the pitfalls and problems of such systems. . . . There is an expression used by some of our ranchers that states, "You can weigh the calf as much as you want, but it won't grow unless you feed it." . . . A lot of what is being done nationally with the design of rath-

Copyright © 1999. All rights reserved.

er heavy-handed, high-stakes accountability systems we believe is nonproductive at best and counterproductive and dead-ended at worst. We agree that we need to clarify the vision, determine our goals and expectations (our standards), determine how we will best measure success once we've defined it, and then ensure that the resources necessary to succeed are committed. Only then does it make sense to put the assessment piece in place. Some will choose to use assessment to drive reform; we are among those who believe that we should use assessment to reflect reform.[40]

---

1. Personal communication, 11 February 1998, from data collected in a current study, sponsored by the Proffitt Foundation, of local administration of state-mandated performance assessments in Pennsylvania, Michigan, and Indiana. Linda Mabry, Katrina Daytner, and Jennifer Aldarondo, *Local Administration of State-Mandated Performance Assessment in Indiana, Michigan, and Pennsylvania: Research Report to the Proffitt Foundation* (Bloomington: Indiana University, 1999).

2. Linda Mabry and Katrina G. Daytner, "State-Mandated Performance Assessment," paper presented at the annual meeting of the American Educational Research Association, Chicago, 1997.

3. Linda Mabry, unpublished data.

4. *Indiana Statewide Testing for Educational Progress: Teacher's Scoring Guide for Grade 10 English/Language Arts* (Monterey, Calif.: CTB/McGraw-Hill, Indiana State Department of Education, 1996), p. 3.

5. Measurement textbooks of all types insist that test constructs, purposes, formats, topics, number of items per topic, and difficulty levels all be predetermined and standardized. For multiple-choice and other selected-response items, the number and characteristics of alternative answers are also to be prespecified. For constructed-response items, the characteristics of answers at each preset performance level are to be prespecified.

6. In the psychometric literature, clear distinctions are made between a test, an assessment, an evaluation, and a measurement. In lay usage, the terms overlap a great deal, and for educators, legislators, and the general public, the psychometric distinctions between them are largely meaningless. I will use them more or less interchangeably here.

7. *Curriculum and Evaluation Standards for School Mathematics* (Reston, Va.: National Council of Teachers of Mathematics, 1989).

8. *Curriculum Standards for Social Studies* (Washington, D.C.: National Council for the Social Studies, 1994); and National Council of Teachers of English and the International Reading Association, *Standards for the English Language Arts* (Urbana, Ill.: NCTE, 1996).

9. Linda Bond, Edward Roeber, and Selina Connealy, *Trends in State Student Assessment Programs: Fall 1997 Data on Statewide Student Assessment Programs* (Washington, D.C.: Council of Chief State School Officers, 1998).

10. J. Collins, "Standards: The States Go Their Own Ways," *Time*, 27 October 1997, p. 75.

11. Robert L. Linn, "Assessment-Based Reform: Challenges to Educational Measurement," first annual William H. Angoff Memorial Lecture, presented at Educational Testing Service, Princeton, N.J., November 1994, pp. 13-17.

12. Linda Darling-Hammond, "Performance-Based Assessment and Educational Equity," *Harvard Educational Review*, vol. 64, 1994, pp. 5-30.

13. Andrew C. Porter, "The Uses and Misuses of Opportunity-to-Learn Standards," *Educational Researcher*, January/February 1995, pp. 21-27.

14. Linda Mabry, *State-Mandated Performance Assessment: Research Report to the Proffitt Foundation* (Bloomington: Indiana University, 1997); and Mabry and Daytner, op. cit.

15. One response has been the newly formed National Center for the Improvement of Educational Assessment, Inc., which, at this writing, is recruiting assessment professionals to assist states in designing and developing assessment programs. Personal communication, Richard Hill, 5 October 1998.

16. *Vermont's Assessment Program* (Montpelier: Vermont Department of Education, 1991).

17. *Kentucky Writing Portfolio, Grade 12 Teacher's Handbook*, 2nd ed. (Frankfort: Kentucky Department of Education, n.d.).

18. Michael Scriven, "The Final Synthesis," *Evaluation Practice*, vol. 15, 1994, pp. 367-82.

19. Linda Mabry, "Naturally Occurring Reliability," paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1995; idem, "Performance Assessment and Inferences of Achievement" (Doctoral dissertation, University of Illinois, Urbana-Champaign, 1995); and Pamela A. Moss, "Can There Be Validity Without Reliability?," *Educational Researcher*, March 1994, pp. 5-12.

20. Wendy M. Yen, "Measuring School Performance: Is 'Percents of Students Reaching Standards' the Most Accurate Statistic?," paper presented at the annual meeting of the American Educational Research Association, Chicago, 1997.

21. See a draft of *Standards for Educational and Psychological Testing* (Washington, D.C.: American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, March 1998), available at www.apa.org/science/standards.html; and Samuel Messick, "Validity," in Robert L. Linn, ed., *Educational Measurement*, 3rd ed. (New York: American Council on Education, Macmillan, 1989), pp. 13-103.

22. George Hillocks, "How State Mandatory Assessment Simplifies Writing Instruction in Illinois and Texas," paper presented at the annual meeting of the American Educational Research Association, Chicago, 1997.

23. Mabry, *State-Mandated Performance Assessment*.

24. Linda Mabry, *Portfolios Plus: A Critical Guide to Alternative Assessment* (Thousand Oaks, Calif.: Corwin Press, 1999).

25. Susan E. Phillips, "Legal Issues in Performance Assessment," *Education Law Quarterly*, vol. 2, 1993, pp. 329-58.

26. In performance assessments in mathematics in Maine and Maryland, aspects of the tests designed to increase reliability so constrained student options that researchers concluded that the change from multiple-choice to performance testing had been superficial rather than conceptual. See William A. Firestone, J. Fairman, and D. Mayrowetz, "The Underwhelming Influence of Testing on Mathematics Teaching in Maine and Maryland," paper presented at the annual meeting of the American Educational Research Association, Chicago, 1997.

27. Edward L. Thorndike, "Measurement in Education," *Teachers College Record*, vol. 22, 1921, p. 378.

28. For a review of these criticisms, see Mabry, *Portfolios Plus*.

29. Mabry, Daytner, and Aldarondo, op. cit.

30. Personal communication, 24 September 1998.

31. Personal communication, 29 April 1998.

32. For another example of testing that deprofessionalized teachers, see Mary Lee Smith, "Put to the Test: The Effects of External Testing on Teachers," *Educational Researcher*, June/July 1991, pp. 8-11.

33. Personal communication, 1 May 1998.

34. Personal communication, 23 September 1998.

35. Hillocks, op. cit.

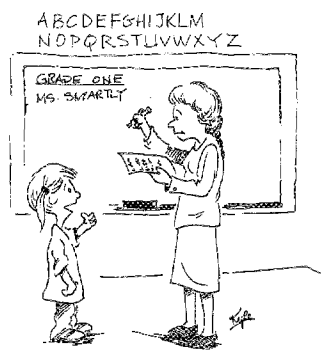
36. Nancy S. Cole, "A Realist's Appraisal of the Prospects of Unifying Instruction and Assessment," in *Proceedings of the Invitational Conference on Assessment in the Service of Learning* (Princeton, N.J.: Educational Testing Service, 1988), pp. 103-16; Lee J. Cronbach, "Five Perspectives on the Validity Argument," in Howard Wainer and H. I. Braun, eds., *Test Validity* (Hillsdale, N.J.: Erlbaum, 1988), pp. 3-17; and Messick, op. cit.

37. Daniel Koretz, *The Reliability of Scores from the 1992 Vermont Portfolio Assessment Program: Interim Report* (Los Angeles: Center for the Study of Evaluation, University of California, Technical Report No. 355, 1992); idem, "Portfolio Assessment: Rhetoric Meets the Reality of Data," paper presented at the annual meeting of the American Educational Research Association, Atlanta, 1993; Daniel Koretz et al., "The Vermont Portfolio Assessment Program: Findings and Implications," *Educational Measurement: Issues and Practice*, vol. 13, 1994, pp. 5-16; and Ronald Hambleton et al., *Technical Review of the Kentucky Instructional Results Information System (KIRIS), 1991-94* (Frankfort: Office of Educational Accountability of the Kentucky General Assembly, 1995).

38. Ernest R. House, "A Framework for Appraising Educational Reforms," *Educational Researcher*, October 1996, pp. 6-14.

39. See, for example, *Education in Indiana: An Overview* (Bloomington: Indiana Education Policy Center, Indiana University, 1994).

40. Jack Gilsdorf, personal communication, 2 January 1997. Nebraska has recently joined the list of states that are in the process of developing standards, often an early step in developing state testing. See Collins, op. cit. K



"I'm checking out. All I need to know I learned last year."

Copyright © 1999. All rights reserved.