0674-1

Practicing What We Preach in Designing Authentic

Designing credible performance tasks and assessments is not easy—but we can improve our efforts by using standards and peer review.

Grant Wiggins

What if a student asked for a good grade merely for handing the paper in? What if student divers and gymnasts were able to judge and score their own performances in meets, and did so based on effort and intent? Naive ideas, of course—yet this is just what happens in schools every day when *faculty* submit new curricular frameworks or design new assessments.

Most faculty products are assessed, if at all, merely on whether we worked hard: Did we hand in a lengthy report, based on lots of discussion? Did we provide students with a test that we happen to like? Only rarely do we demand formal self- or peer-assessment of our design work, against standards and criteria. This not only leads to less rigorous reports and designs but also seems a bit hypocritical: We ask students to do this all the time. We need to better practice what we preach.

But how do we ensure that ongoing design and reform work is more rigorous and credible? At the Center on Learning, Assessment, and School Structure (CLASS) in Princeton, New Jersey, we use design standards and a workable peer review process for critiquing and improving all proposed new curricular frameworks, tests, and performance assessments. At the heart of the work is making adult work standards-based, not process-based or merely guided by good intentions. Using such standards can go a long way in helping parents, students, and the community have faith in locally designed systems.

Standard-Based vs. Process-Based Reform Work

Many new curriculum frameworks and assessment systems produce a significant (and often understandable) backlash. A major reason is that the work is typically produced without reference to specific standards for the proposals and final product.

Think of a typical districtwide curriculum reform project. Twelve teachers and supervisors hold meetings all school year to develop a new mathematics curriculum. Their work culminates in a report produced over a three-week period in the summer, at district behest and with district financial support, resulting in a new local mathematics curriculum framework. They follow a time-tested *process* of scanning national reports, searching for consensus about themes and topics and logical progressions, and summarizing their findings and recommendations. But against what standards is their *product* (as opposed to their process) to be judged? The usual answer is: no legitimate standards at all, other than the implicit one that when the authors deem their work finished, the report is complete.

By contrast, what if all report-writers had to answer these questions: Is the report useful to readers? Does it engage and inform the readers? Does it anticipate the reactions of its critics? Does it meet professional standards of curriculum design or measurement? Does it meet the purposes laid out in a charge to the committee? Most important: *Did the writers regularly self-assess and revise their work in progress against such criteria and standards? Did they regularly seek feedback from faculty affected en route?*—the same writing process questions we properly put to students. Their report would have far greater impact if they addressed such questions. By contrast,

© Suse Fitzhugh

18

0674-2

Assessments

with no self-assessment and self-adjustment along the way, the work is predictably ineffective in getting other faculty to change practice or in helping skeptical parents understand the need to do so.

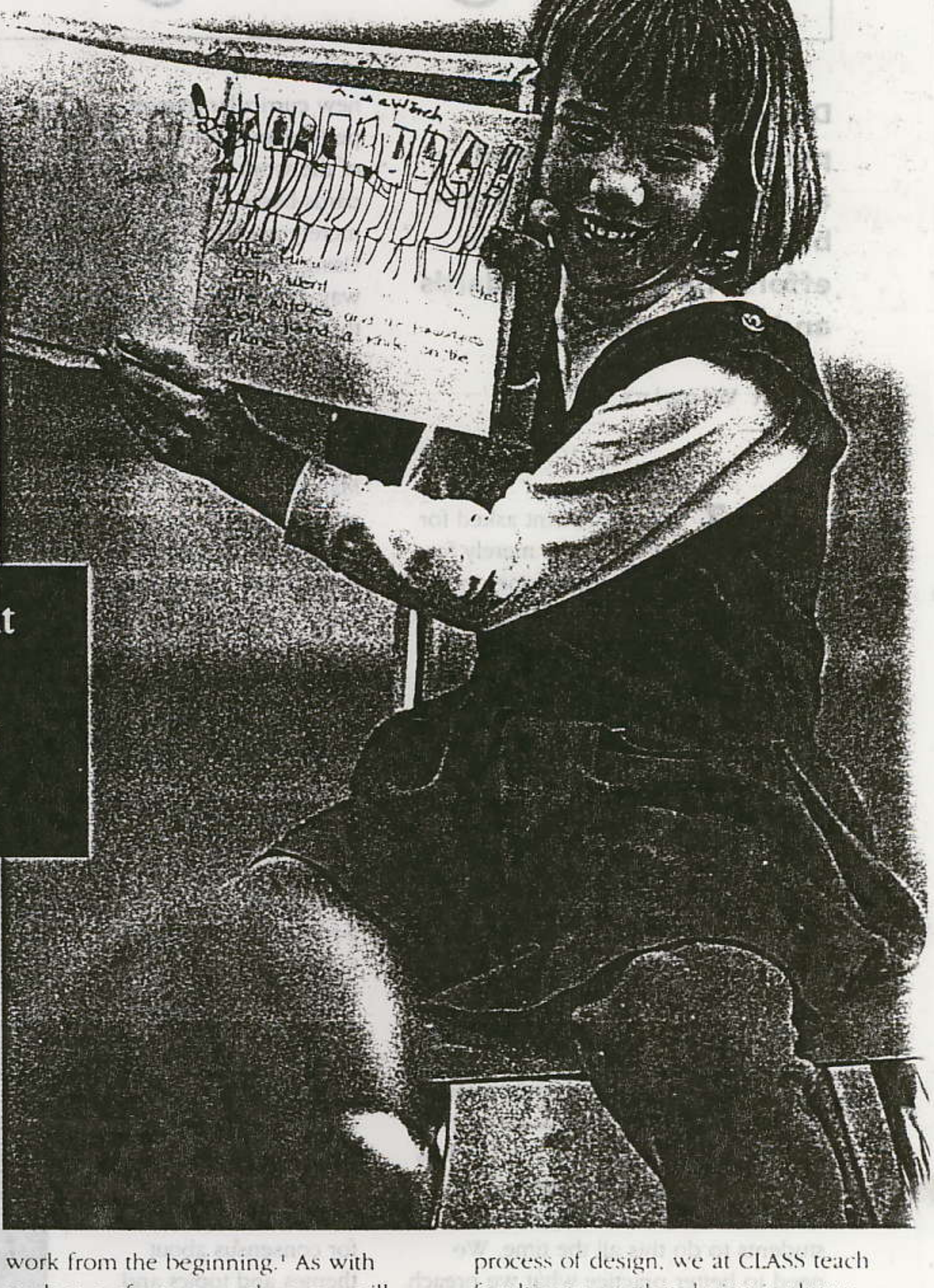
Similarly with new assessments. Almost every teacher designs tests under the most naive premise: "If I designed it and gave it, it must be valid and reliable." Yet we know from research, our own observations, and the process of peer review that few teacher-designed tests and assessments meet the most basic standards for technical credibility, intellectual defensibility, coherence with system goals, and fairness to students.

The purpose of assessment is to find out what each student is able to do, with knowledge, in context.

When we practice what we preach about self-assessment and adjustment against standards, we can *ensure* more rigorous and effective local teacher products, greater collegiality, and better student performance.

In standards-based reform projects, in short, we must seek a disinterested review of products against standards all along the way—not just follow a process in the hope that our work turns out well. The challenge for school reformers is to ensure that their work has *impact*, like any other performance. Desired effects must be *designed in;* they must inform all our work from the beginning.[1] As with student performances, then, we will meet standards only by "backwards design"—making self-assessment and peer review against performance standards central to the process of writing and revision—*before* it is too late.

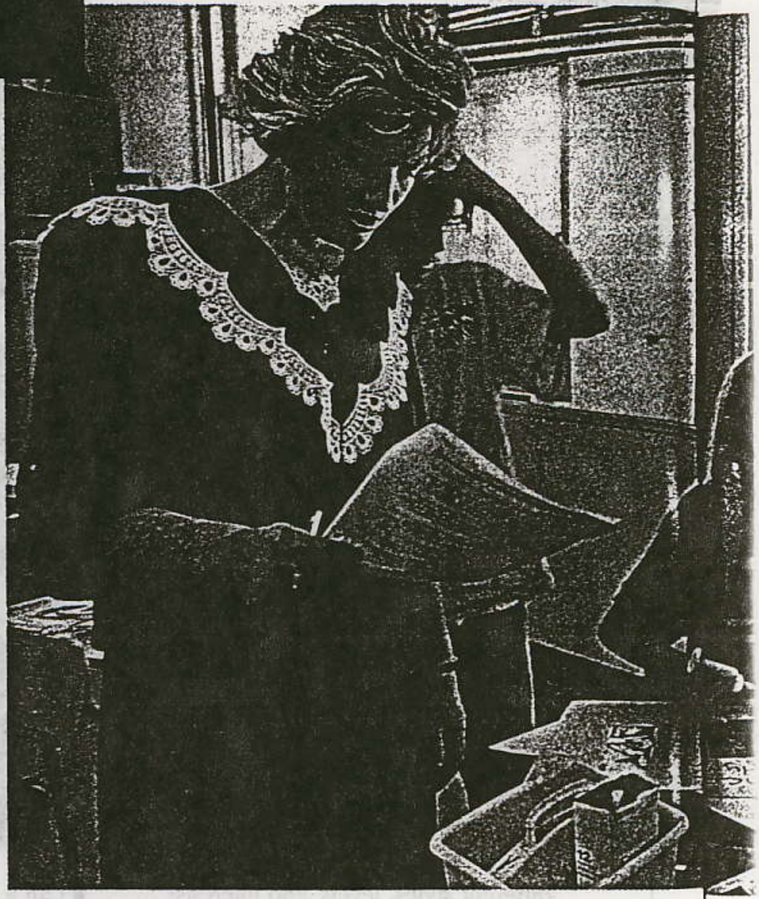
Rather than teaching a lock-step process of design, we at CLASS teach faculties to see that design is always *iterative*. We constantly rethink our designs, using feedback based on clear design standards. We will likely never revisit our original designs if we lack powerful criteria and a review process with the implicit obligation to critique



© Susie Fitzhugh

0674-3

Complex performance tasks focus on understanding as an educational goal, as opposed to mere textbook knowledge.



all work against the criteria. We are often satisfied with (and misled by) our effort and good intentions.

Assessment Design Standards

Standards-based reform work begins with clear standards for eventual products. At CLASS, we instruct faculties involved in performance-based assessment reform in the use of a design template, a design process, and a self-assessment and peer review process based on ultimate-product standards. In addition, we work with leaders to make such standards-based design work more routine in and central to :al faculty life (linked to job descriptions, department meetings, and performance appraisal systems, as well as individual and team design work). The template is also the database structure for assessment tasks and rubrics on our World Wide Web site, http://www.classnj.org.

The standards guide all design decisions. The three main criteria for judging emerging tasks are *credibility, user-friendliness,* and *feasibility.* The standards are fixed by specific models that serve to anchor the self-assessment and peer review process (just as in the assessment of student writing). Each criterion is broken down further into subcriteria: Under credibility, for example, the designer (in self-assessing) and the peers (in peer reviewing) ask such questions as:

- Does it measure what it says it measures? Is this a valid assessment of the intended achievement?
- Are the scoring criteria and rubrics clear, descriptive, and explicitly related district goals and standards?
- Is the scoring system based on genuine standards and criteria, derived from analysis of credible models?
- Does the task require a sophisticated understanding of required content?
- Does the task require a high degree of intellectual skill and performance quality?
- Does the task simulate or replicate authentic, messy, real-world challenges, contexts, and constraints faced by adult professionals, consumers, or citizens?
- Does the scoring system enable a reliable yet adequately fine discrimination of degrees of work quality?
- Is the task worthy of the time and energy required to complete it?
- Is the task challenging—an appropriate stretch for students?

Naturally, in parallel to what we ask of students, there are rubrics for self-and peer-assessment of these questions.

Anticipating Key Design Difficulties

We ask designers to pay particular attention to three crucial, ever-present problems in local assessment design: whether a sophisticated understanding of core content is required by the task, whether the criteria and rubrics used are authentic and appropriate for such a task and target, and whether the tasks really measure the targeted achievement. This last problem can be stated as a single injunction that must be constantly invoked: Beware the temptation of confusing a neat instructional activity with an appropriate performance task.

1. Validity in design. Validity is essential. The purpose of assessment is to find out what each student is able to do, with knowledge, in context. But we must sample from a large domain. In asking students to do a *few* tasks well, we believe we are on solid ground because we view the tasks as apt—at the heart of the subject, and able to yield more general inferences about achievement in a subject.

When we worry about validity in design, we are thinking backwards from the evidence we need. The task must yield the right kind of information and must enable us to elicit and observe the most salient performance, given the (more general) achievements we seek to measure.

In instruction, our worries are different. We typically try to develop activities that give rise to an educational experience and ask questions that differ from those that apply to assessment design: Will the students be engaged? Will we accommodate

0674-4

Complex performance tasks focus on understanding as an educational goal, as opposed to mere textbook knowledge.

© Nita Winter

FIGURE 1

What Does Understanding Mean?

Complete the following sentence to help construct an authentic, credible performance assessment in any subject matter:

The students *really* understand (the idea, issue, theory, event being assessed) only when they can...

- provide credible theories, models, or interpretations to explain ...
- avoid such common misunderstandings as ...
- make such fine, subtle distinctions as ...
- effectively interpret such ambiguous situations or language as ...
- explain the value or importance of ...
- critique ...
- see the plausibility of the "odd" view that ...
- empathize with ...
- critically question the commonly held view that ...
- invent ...
- recognize the prejudice within that ...
- question such strong personal but unexamined beliefs as ...
- accurately self-assess ...

different styles, levels, and interests? Will the activity give rise to thinking and learning at the heart of my goal for the unit? Such questions are essential to teaching, but unlikely to ensure that we will have adequate assessment evidence for *each* student when the activity is over.

Easy to say, but what to do? That's where the peer review process comes in. We are now forced to *justify* our design in a nonconfrontational way. In peer review, we often discover that the design does not yet work as a sound assessment. (Eventually, our self-assessment becomes so skilled that we can foresee these kinds of problems without much peer review.)

These are the questions we use in peer review for validity:

- Does the task evoke the right kind of evidence, given the target? Does the task evoke sufficient evidence?
- Can a student master the task for the "right" reasons only? Or does the task unwittingly assess for a different outcome than intended by the designer?

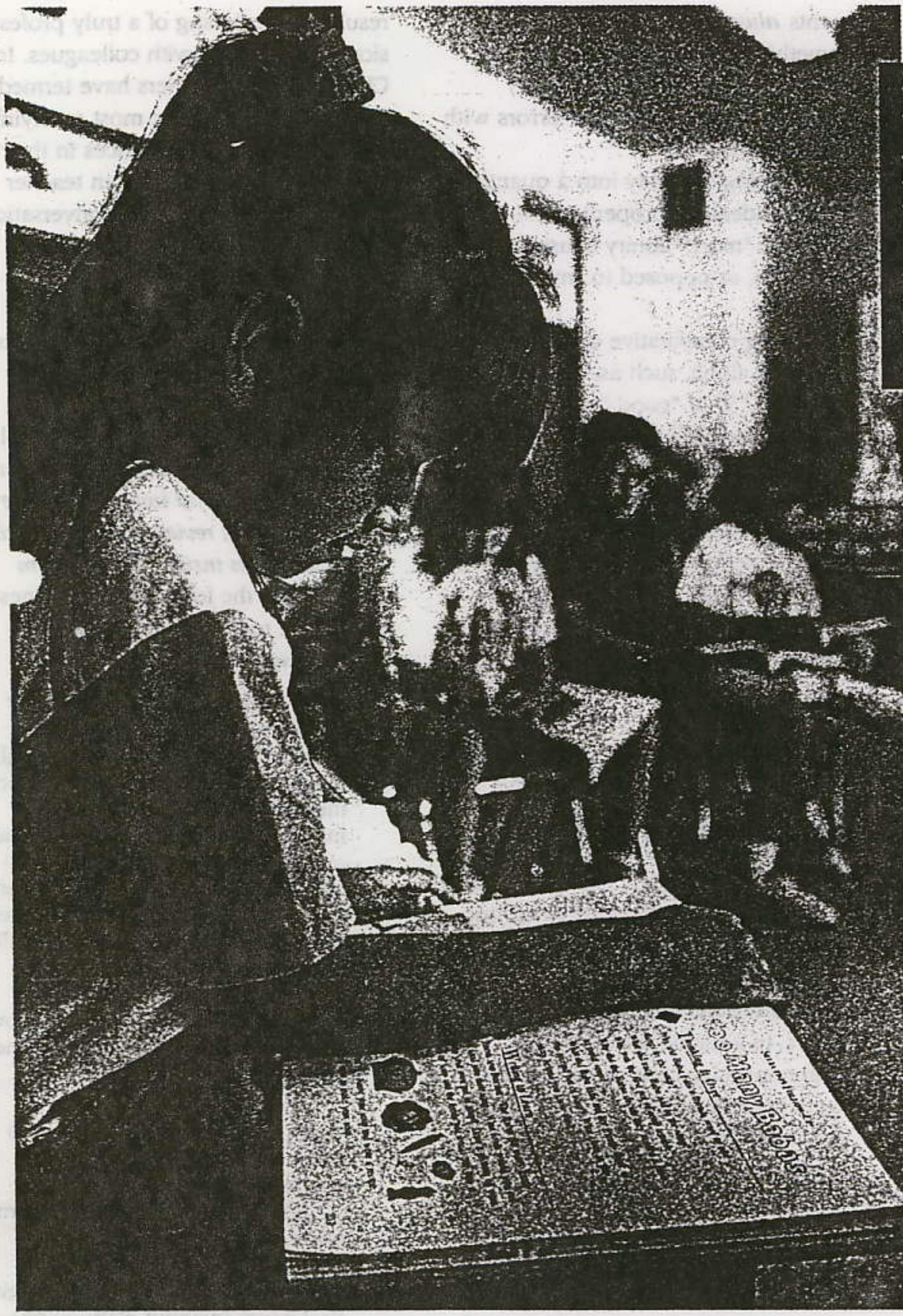
Yes, it measures what it's supposed to if the task can only be done well if students are in control of the key achievements.

No, it doesn't measure what it should if students (1) can perform the task well without achieving the intended result or (2) fail to perform the task well for inappropriate reasons, that is, abilities or knowledge unrelated to the target.

- Are the criteria apt? That is, given the achievements to be assessed and the nature of the task, are these the right traits of performance to assess and the right descriptions of differences in work quality?
- Is the weighting of the different criteria appropriate, given the nature and purpose of such performance?
- Do the scoring rubrics discriminate levels of quality appropriately and not arbitrarily?
- Does the task imply a rich and appropriate understanding of the intended target? Or is the task implicitly based on a questionable or inappropriate definition of the achievement?
- Do the rubrics honor the criteria and achievement? Or are they implicitly based on questionable or inappropriate definitions of exemplary performance?

2. Assessment for understanding. Because any complex performance tends to focus on fairly general academic skills, performance tasks often unwittingly lack sufficient intellectual rigor and credibility.[2] Many tasks simply reveal whether students can "communicate" or "problem solve"—

0674-5



© Jonathan Meyers

Peer review can yield a profound result: the beginning of a truly professional relationship with colleagues.

d often allow great leeway in subject-matter content. Consider the specific knowledge required to perform these complex tasks developed by teachers in North Carolina:

Birds and Soldiers. Wildlife officials and politicians are at odds because of the rare red-cockade woodpecker on the Fort Bragg military base. Fort Bragg officials have to limit military training exercises because of the protection required for the birds under the Endangered Species Act. The Act states that an endangered bird's environment cannot be tampered with. Almost half the known red-cockade woodpecker population is located on the base. Your task is to propose a workable solution to the problem, based on a careful review of the military's needs and the relevant law. You will write a report and make a speech to a simulated EPA review board.

Federation/Confederation. This task involves three parts: a) the student is asked to assume the role of a resident of North Carolina on the eve of secession and deliver a speech from that person's perspective on whether or not North Carolina should secede from the Union, b) the student then synthesizes the points from all speeches given and writes a letter to the editor of the local newspaper reflecting this person's re-examined point of view, and c) writes a reflective piece in the person's journal, 15 years later, re-examining the wisdom of the earlier stands.

It's Your Choice: Health Insurance. Co-payment? Pretreatment estimate? Deductible? Is health care language a foreign language for you? Students take on the role of a financial analyst and must communicate to each of three different families, in a convincing manner, the best choice of coverage for their needs and budget.

These tasks focus on *understanding* as an educational goal, as opposed to mere textbook knowledge. We at CLASS have developed a complex schema for teaching and assessing

0674-6

understanding, drawing not only on our own research of the past decade but also the fine work of Howard Gardner (1992), David Perkins (1992), and their Project Zero colleagues. As we see it, to assess for understanding means to assess for five related capacities: sophistication of explanations and interpretations; insight gained from perspective; empathy; contextual know-how in knowledge application; and self-knowledge based on knowing our talents, limits, and prejudices.

What is the evidence we need to gather? At CLASS, we use the exercise in Figure 1 as a reminder. As a prompt, we ask teachers to brainstorm ways to complete the sentence stem that reads, "The students understand the idea only when they can..." Then we integrate the brainstormed ideas by building a rubric of sophisticated understanding on a novice-to-expert continuum. For example, take key events in history: What is a novice versus a sophisticated understanding of the Civil War? What sorts of judgments and discriminations is an expert likely to make that a novice student is unlikely to make? Such questions force us to predict how students are likely to perform.

The most exciting effect of this exercise is to realize that we must be able to predict students' inevitable misunderstandings. Of all the assessment strategies we have used, this is the one that causes the most "Aha!" responses. To teach and assess *mindful of misunderstanding* requires not only rubrics for levels of understanding and misunderstanding, but a new perspective on teaching: If you can now predict student misunderstandings, what are you doing to avoid or aggressively compensate for them in your curriculum and instruction?[3]

3. Critique and revision of rubrics and criteria. The designer of assessments *always* has a blind spot about something. Peers can discover and help to remedy oversights. The following represent typical errors with most rubrics:

- Turning a quality into a quantity. Thus, students improperly get a higher score for "more" library sources or footnotes, as opposed to "more apt" sources.
- Using comparative or evaluative language alone, such as "6" or "excellent" and "5" or "good," and so forth, when observable traits of performance are more meaningful.
- A lack of continuity in the "distance" between score points. Thus, in the descriptors for a 6, 5, and 4, the differences may be slight. Suddenly, a 3 is just awful and not passing, so the score points are bunched at one end and spread out at another, causing misleading results.

Other pitfalls to watch for include combining traits, such as "creative" and "organized," in the same descriptor, and confusing a criterion with its indicators. For example, "asking questions" is an indicator of *good listening,* but silence in church doesn't mean that people aren't listening. Inappropriate questions don't indicate good listening, either. In addition, most rubrics overemphasize content and form of the work and underemphasize or ignore the *impact* of performance—criteria at the heart of what we mean by "performance."

Most of us make these mistakes when we begin writing rubrics. Peer review, based on design standards, ensures that rubrics are debugged of common mistakes.

Peer Review

Besides improving the process of developing performance assessments, peer review can yield a profound result: the beginning of a truly professional relationship with colleagues. In CLASS projects, teachers have termed peer review one of the most satisfying (if initially scary) experiences in their careers. As a 32-year veteran teacher put it, "This is the kind of conversation I entered the profession to have, yet never had. I'm rejuvenated. I'm optimistic."

Peer reviewers serve as consultants to the designer, not glib judges. The process itself is evaluated against a basic criterion in support of that goal: *The designer must feel that the design was understood and improved by the process, and the reviewers must feel that the process was insightful and team building.* As the following guidelines reveal, the reviewers give specific, focused, and useful feedback:

> *Stage 1: Peers review task without designer present.*[4] The designer states issues he or she wishes highlighted, self-assesses (optional), and then leaves. The peers read the materials, referring to the *assessment design criteria*. Working individually, the peers summarize the work's strengths and weaknesses and then report to the group. The group fills out a sheet summarizing the key feedback and guidance, thus rehearsing the oral report to follow. Reviewers rate the task against the task rubric, if appropriate.
>
> *Stage 2: Peers discuss review with designer.* Appointing a timekeeper/facilitator is crucial. The facilitator's job is to gently but firmly ensure that the designer listens (instead of defending). First, the designer clarifies technical or logistical issues (without elaboration)—*the design must stand by itself as much as possible.* Second, the peers give oral feedback and guidance. Third, the group and the designer discuss the feedback; the designer takes notes and asks questions. Finally, the group decides what

0674-7

issues should be presented to the faculty as a whole—lessons learned and problems evoked.

Criteria for peer review:

1. The core of the discussion involves considering: To what extent is the "targeted achievement" well assessed? To what extent do the task and rubric meet the design criteria? What would make the assessment more valid, reliable, authentic, engaging, rigorous, fair, and feasible?

2. The reviewers should be friendly, honest consultants. The designer's intent should be treated as a given (unless the unit's goal and means are unclear or lack rigor). *The aim is to improve the designer's idea, not substitute it with the reviewers' aesthetic judgments, intellectual priorities, or pet designs.*

3. The designer asks for focused feedback in relation to specific design criteria, goals, or problems.

4. The designer's job in the second session is primarily to listen, not explain, defend, or justify design decisions.

5. The reviewers' job is first to give useful *feedback* (did the effect match the intent?), and only then, useful *guidance.*

Note that we distinguish here between feedback and guidance. The best feedback is highly specific and descriptive of how the performance met standards. Recall how often a music teacher or tennis coach provides a steady flow of feedback (Wiggins 1993). Feedback is *not* praise and 'ame or mere encouragement. Try ecoming better at any performance if all you hear is "Nice effort!" or "You can do better" or "We didn't like it." Whatever the role or value of praise and dislike, they are not feedback: The information provided does not help you improve. In feedback and guidance, *what matters is judging the design against criteria related to sound assessment.* Peer reviewers are free to offer concrete guidance—suggestions on how the design might be improved—assuming the designer grasps and accepts the feedback.[5]

Assessment System Criteria

Beyond reviewing specific performance tasks and rubrics, we need to evaluate entire assessment systems. For such systemic assessments, a more complex set of criteria includes credibility, technical soundness, usefulness, honesty, intellectual rigor, and fairness (Wiggins 1996).

Again, a key to credibility is *disinterested* judging—using known and intellectually defensible tasks and criteria—whether we are talking about student or faculty work. A psychometrician may well find a local assessment system not up to a rigid technical standard; but such a system can still be credible and effective within the real-world constraints of school time, talent, and budgets.

Credibility is a concern of the whole school community. We need other feedback—not just from peer reviewers, teacher-designers, or psychometricians, but from parents, school boards, college admissions officers, and legislators. Alas, what one group finds credible, another often doesn't. Clients for our information have differing needs and interests in the data; if we fail to consider these clients, our local assessment systems may be inadequate and provincial. But if we improperly mimic large-scale, audit testing methods in an effort to meet psychometric standards for local assessment design, we often develop assessment systems that are neither authentic nor effective as feedback.

Consider the possible customers for the assessment information, to determine whether both the task and the reporting of results are apt and adequate.

Peer review should always consider the possible customers for the assessment information, to determine whether both the task and the reporting of results are apt and adequate (Wiggins 1996). The primary customer is always the student.

Principles Underlying the Standards and Criteria

When proposing standards and criteria for performance assessments, we need to remember—and clearly state—the underlying values of our proposals. Assessment is not merely a blind set of techniques, after all, but a means to some valued end. Effective and appropriate school assessment is based on five principles:

1. Reform focuses on the purpose, not merely the techniques, of assessment. Too many reform projects tamper with the technology of assessment without reconnecting with the purposes of assessment. Assessment must recapture essential educational aims: to help the student learn and to help the teacher instruct. All other needs, such as accountability testing and program evaluation, come second. Merely shifting from multiple-choice questions to performance testing changes nothing if we still rely on the rituals of year-

0674-8

end, secure, one-shot testing.

2. *Students and teachers are entitled to a more instructional and user-friendly assessment system than provided by current systems and psychometric criteria.* A deliberately instructional assessment makes sure that tests enlighten students about real-world intellectual tasks, criteria, context, and standards; and such an assessment is built to ensure user-friendly, powerful feedback. Conventional tests often prevent students from fully understanding and meeting their intellectual obligations. And teachers are entitled to an accountability system that facilitates better teaching.

3. *Assessment is central, not peripheral, to instruction.* We must design curriculums backwards from complex and exemplary challenges. A performance-based system integrates curriculum and assessment design, thereby making the sequence of work more coherent and meaningful from the learner's point of view.

4. *Authentic tasks must anchor the assessment process, so that typical test questions play a properly subordinate role.* Students must see what adults really do with their knowledge; and all students must learn what athletes already know—that performance is more than just the drill work that develops discrete knowledge and skill. Genuine tasks demand challenges that require good judgment, adaptiveness, and the habits of mind—such as craftsmanship and tolerance for ambiguity—never tested by simplistic test items.

5. *In assessment, local is better.* Site-level assessments must be of higher intellectual quality—more tightly linked to instruction—than superficial standardized tests can ever be. No externally run assessment can build the kind of local capacity for and interest in high-quality assessment at the heart of all genuine local improvement. But local assessment must be credible—and that means inviting disinterested assessment by people other than the student's teachers, and including oversight of the entire assessment design and implementation system (for case studies in assessment reform, see CLASS 1996).

By keeping these principles in mind, we can continually improve our reform work. Process-driven improvement efforts can become rigid and noncreative; we resort to following the letter of the law only. The real power of standards-based reform is that we are free to innovate and divert from process—if we see a better way to approach the standards and better honor our principles. Thus, our reform efforts, not just our designs, also demand constant self-assessment and self-adjustment, based on comparing emerging work against our principles.[6]

Professionalism depends on standards-based work and peer review. Despite the long-standing habits of schools where teachers are left alone to design assessments, we believe that such practices are counterproductive to both local credibility and professional development. Every school and district ought to require peer review of major assessments, based on sound and agreed-upon standards and criteria of design and use. ■

[1] For student performance tasks, too, rubric and task writers should emphasize impact-related criteria so that students know the purpose of the task. Thus, instead of just scoring for organization, clarity, and accuracy in essay writing, we should include criteria related to how persuasive and engaging the piece is.

[2] Bob Marzano believes that performance assessment is ill-suited for assessing understanding of subject matter. I disagree: Intellectual understanding is demonstrated by doing well at certain types of performance, but designing such tasks is indeed difficult.

[3] A full development of this schema of understanding will appear in 1997 in a new ASCD book and training program, co-authored by Jay McTighe and myself, and tentatively titled *Understanding by Design.*

[4] Some may wonder about the utility or ethics of discussing the work in the designer's absence. We have found that this first stage gives the peers freedom to express vague concerns and complete criticisms. When the designer is always present, we find that the session bogs down as the designer justifies and explains all decisions.

[5] Video and print material on the peer review process is available from CLASS.

[6] Fairtest (1995) has developed standards and indicators for assessment processes and systems. Contact Fairtest at National Center for Fair & Open Testing, 342 Broadway, Cambridge, MA 02139. Phone: (617) 864-4810; fax: (617) 497-2224; e-mail: FairTest@aol.com.

References

Center on Learning, Assessment, and School Structure (CLASS). (1996). *Measuring What Matters: The Case for Assessment Reform* (video). Princeton, N.J.: CLASS.

Fairtest: National Center for Fair and Open Testing. (1995). *Principles and Indicators for Student Assessment Systems.* Cambridge, Mass.: Fairtest.

Gardner, H. (1992) *The Unschooled Mind.* New York: Basic Books.

Perkins, D. (1992). *Smart Schools: Better Thinking and Learning in Every Child.* New York: Free Press.

Wiggins, G. (1993). *Assessing Student Performance: Exploring the Purpose and Limits of Testing.* San Francisco: Jossey-Bass.

Wiggins, G. (1996). "Honesty and Fairness: Toward Better Grading and Reporting," in *Communicating Student Learning,* edited by T. Guskey. 1996 ASCD Yearbook. Alexandria, Va.: ASCD.

Grant Wiggins is President of the Center on Learning, Assessment, and School Structure (CLASS), 648 The Great Road, Princeton, NJ 08540. He can be reached by e-mail at gpw@classnj.org.