

Withdrawal/Redaction Sheet

Clinton Library

DOCUMENT NO. AND TYPE	SUBJECT/TITLE	DATE	RESTRICTION
--------------------------	---------------	------	-------------

001. Fax	Post-it Fax Note re: phone number (partial) (1 page)	10/03/1999	P6/b(6)
----------	--	------------	---------

COLLECTION:

Clinton Presidential Records
Domestic Policy Council
Andrew Rotherham (Education)
OA/Box Number: 21293

FOLDER TITLE:

Teacher Evaluations

2011-0103-S

rc142

RESTRICTION CODES

Presidential Records Act - [44 U.S.C. 2204(a)]

Freedom of Information Act - [5 U.S.C. 552(b)]

- P1 National Security Classified Information [(a)(1) of the PRA]
- P2 Relating to the appointment to Federal office [(a)(2) of the PRA]
- P3 Release would violate a Federal statute [(a)(3) of the PRA]
- P4 Release would disclose trade secrets or confidential commercial or financial information [(a)(4) of the PRA]
- P5 Release would disclose confidential advice between the President and his advisors, or between such advisors [(a)(5) of the PRA]
- P6 Release would constitute a clearly unwarranted invasion of personal privacy [(a)(6) of the PRA]

C. Closed in accordance with restrictions contained in donor's deed of gift.

PRM. Personal record misfile defined in accordance with 44 U.S.C. 2201(3).

RR. Document will be reviewed upon request.

- b(1) National security classified information [(b)(1) of the FOIA]
- b(2) Release would disclose internal personnel rules and practices of an agency [(b)(2) of the FOIA]
- b(3) Release would violate a Federal statute [(b)(3) of the FOIA]
- b(4) Release would disclose trade secrets or confidential or financial information [(b)(4) of the FOIA]
- b(6) Release would constitute a clearly unwarranted invasion of personal privacy [(b)(6) of the FOIA]
- b(7) Release would disclose information compiled for law enforcement purposes [(b)(7) of the FOIA]
- b(8) Release would disclose information concerning the regulation of financial institutions [(b)(8) of the FOIA]
- b(9) Release would disclose geological or geophysical information concerning wells [(b)(9) of the FOIA]

Withdrawal/Redaction Marker

Clinton Library

DOCUMENT NO. AND TYPE	SUBJECT/TITLE	DATE	RESTRICTION
001. Fax	Post-it Fax Note re: phone number (partial) (1 page)	10/03/1999	P6/b(6)

COLLECTION:

Clinton Presidential Records
Domestic Policy Council
Andrew Rotherham (Education)
OA/Box Number: 21293

FOLDER TITLE:

Teacher Evaluations

2011-0103-S

rc142

RESTRICTION CODES

Presidential Records Act - [44 U.S.C. 2204(a)]

- P1 National Security Classified Information [(a)(1) of the PRA]
- P2 Relating to the appointment to Federal office [(a)(2) of the PRA]
- P3 Release would violate a Federal statute [(a)(3) of the PRA]
- P4 Release would disclose trade secrets or confidential commercial or financial information [(a)(4) of the PRA]
- P5 Release would disclose confidential advice between the President and his advisors, or between such advisors [(a)(5) of the PRA]
- P6 Release would constitute a clearly unwarranted invasion of personal privacy [(a)(6) of the PRA]

C. Closed in accordance with restrictions contained in donor's deed of gift.

PRM. Personal record misfile defined in accordance with 44 U.S.C. 2201(3).

RR. Document will be reviewed upon request.

Freedom of Information Act - [5 U.S.C. 552(b)]

- b(1) National security classified information [(b)(1) of the FOIA]
- b(2) Release would disclose internal personnel rules and practices of an agency [(b)(2) of the FOIA]
- b(3) Release would violate a Federal statute [(b)(3) of the FOIA]
- b(4) Release would disclose trade secrets or confidential or financial information [(b)(4) of the FOIA]
- b(6) Release would constitute a clearly unwarranted invasion of personal privacy [(b)(6) of the FOIA]
- b(7) Release would disclose information compiled for law enforcement purposes [(b)(7) of the FOIA]
- b(8) Release would disclose information concerning the regulation of financial institutions [(b)(8) of the FOIA]
- b(9) Release would disclose geological or geophysical information concerning wells [(b)(9) of the FOIA]

Post-it® Fax Note	7671	Date	# of pages
To JB Buxton	From B. BERRY		
Co./Dept.	Co.		
Phone #	Phone #		
Fax # 202-456-7328	Fax #		

[001]

CHAPTER 23

Toward What End?

The Evaluation of Student Learning for the Improvement of Teaching

LINDA DARLING-HAMMOND

There are all kinds of technical quarrels that can be raised about student achievement measures and their use in various indexes, regression models, or teacher evaluation systems. However, the bottom-line questions for any assessment system that purports to measure the quality of schools or teaching are two:

1. Is it real? Is the system really measuring the quality of schooling or teaching? Or is it measuring something else, such as changes in student population or artifacts of the assessment methods?
2. What are the effects? Does it improve teaching and learning? Conversely, does it do any harm?

In this response, I examine the systems reviewed in this book through the lens of these two questions. Unless the answers to both of them are positive, researchers may be amusing themselves or satisfying the whims of policymakers, but they are doing little to help teachers and students improve their collective work. I conclude by suggesting what we might ideally want to see in an assessment system that can both evaluate the quality of teaching and support educators in their efforts to improve.

The Kentucky Instructional Results Information System

Kentucky's experience illustrates how a good idea for improving curriculum and student assessment can be undermined by a bad idea for "accountability"—one that fails to measure the quality of schooling and undermines teaching in the process.

In many respects, Kentucky's efforts to create and implement a new performance-based assessment system in a short period of time might be viewed as nothing short of heroic. Evaluations of the student assessments themselves point to work that is needed to improve scoring reliabilities, but many analysts also note that the assessments have begun to influence instruction in positive ways, especially in encouraging much more extensive and higher quality student writing (Appalachia Educational Laboratory, [AEL], 1996; Whitford & Jones, in press; also see Stufflebeam's contribution to this book [Chapter 20]). My own review suggests that tasks on the early Kentucky instruments were more authentic than those on traditional standardized tests; most were well constructed and provided useful indicators of student learning.

The Kentucky Department of Education was criticized by some for eliminating multiple-choice items from the tests and for continuing to use portfolios despite low reliabilities in scoring. (Recently, multiple-choice items have been reinstated and the mathematics portfolio has been shelved.) Certainly, the concern for adequate reliability in a high-stakes system is justified. However, as performance assessments evolve, it is increasingly possible to score them reliably, and Kentucky eventually should be able to benefit from the accumulated wisdom of other states. In recent work in New York, for example, we have obtained interrater reliabilities for performance assessments in reading, writing, and mathematics consistently above .85 and often above .90 (Pechione, Falk, & Darling-Hammond, in press), even though these include extended tasks and no multiple-choice items.

Even where such results have not yet been obtained, however, I believe the Department of Education's original decisions can be defended on the grounds of both consequential and construct validity. The consequential validity case is strengthened as reports of significantly improved instruction accompany the use of performance assessments and portfolios in states as far-flung as Vermont, New York,

and California as well as Kentucky (AEL, 1996; Darling-Hammond & Falk, in press; Koretz, Stetcher, & Deibert, 1992; Murnane & Levy, 1996).

In terms of construct validity, it is time to demand serious answers to the question of what multiple-choice responses really measure, when they can be answered with guesswork rather than knowledge, and when they give no indication of the ability to apply information in a performance context. The critique of multiple-choice tests for their failure to measure complex performance skills is well known (see, e.g., Resnick, 1987). However, even straightforward factual knowledge may be poorly measured by such items. One study that asked students to reply to the same set of questions posed in multiple-choice and short-answer formats found that scores were 43% higher when students were able to pick answers out of a list than when they had to produce the answers themselves. Furthermore, many of the constructed responses were so far off the mark that it was clear the "correct" multiple-choice answers measured anything but genuine learning (Shea, 1993).

We may be asking the wrong questions about the relationship between assessment methods and stakes. Rather than limiting the range of assessments because of the stakes attached to them, perhaps we should be limiting the range of stakes to meet the demands of educationally useful assessments. If assessments are created that produce improvements in teaching and learning, then we should ask whether stakes that undermine these influences are warranted and defensible.

This brings me to the fly in the Kentucky ointment: its poorly conceived accountability system. As reviewers in this book and elsewhere have noted, there are many technical problems with the accountability index: the arbitrary reduction of many dimensions to a single score; unresolved questions about its reliability; the arbitrary setting of expected improvement levels from schools in continuous, linear fashion; the failure to take into account differences in student populations and community conditions; and the failure to allow for cohort analyses of the same students over time, a particularly problematic shortcoming.

However, the most egregious limitations of the accountability system are the fact that it does not measure school quality in a meaningful way and it undermines high-quality teaching in a number of ways. The use of cross-sectional rather than cohort analyses of per-

formance means that annual scores are as likely to be an artifact of changes in the population of students taking the test as in the quality of teaching they experience. For example, one Kentucky elementary school of 200 students, a reputedly excellent school visited by many educators seeking to understand good practice, found itself on the "merit school" list one year and, with a small change in the students enrolled, on the "deficient" list the following year without having undergone any change in school practices. Five out of nine schools on the "deficient" list this year were "in rewards" last year (Whitford & Jones, in press). This common occurrence destroys the credibility of the so-called accountability system, which clearly is not measuring school quality at all.

As Stufflebeam's analysis suggests, studies have found that educators in many schools, particularly small schools and those with large numbers of disadvantaged students or high turnover rates, feel victimized by the accountability system because it does not measure the actual progress of their students and evaluates them unfairly (AEL, 1996; Whitford & Jones, in press). Furthermore, the rewards and sanctions scheme assumes that the real problems of schools are caused by teachers who are holding out on their students; that is, that they will only teach well out of greed or fear. According to reports from the field, this offends many educators while it heightens their anxiety, which they pass on to their students by applying negative pressure, creating rewards and sanctions for students, and narrowing their teaching to test-related drills that are becoming narrower as the tests become less authentic and performance oriented (Whitford & Jones, in press). Some evidence suggests that, because of these pressures, increased scores are a result of better test taking rather than genuinely improved learning (Hambleton et al., 1995).

The accountability system never sought to improve learning directly by investing in teachers' abilities to teach more effectively. Furthermore, it undermines the teacher learning and enhanced teaching quality that might result from more instructionally valid and powerful assessments, which are now being removed from the system because they are not scored reliably enough to support high-stakes decisions. Finally, the high-stakes nature of the assessments has minimized teacher involvement in the assessment development and scoring process, which would increase knowledge about how to teach. Costly external contractor scoring eliminates the potential learning about curriculum, assessment, and teaching that teachers

could get from being involved in moderated scoring of student work. This may account for the finding that teachers have not integrated assessment into their instructional practices (see Chapter 20), and it may lead to the ironic outcome that the nation's most ambitious plan to overhaul curriculum and teaching was short-circuited by the very policy that was supposed to make schools pay attention to it.

The Dallas Value-Added Accountability System

Unlike Kentucky, the Dallas accountability system (which was proposed but not ultimately implemented) at least strives to evaluate value-added learning by measuring individual student test scores using prior scores as controls. This intention confronts a major Achilles' heel in the approach used in many states and districts, although, as Thum and Bryk note in their contribution to this book (Chapter 9), a gain score analysis would be preferable to the use of residuals that introduce error and reduce expectations for many students.

However, the proposed Dallas system seemed to have little to do with either the appropriate evaluation of teaching or its improvement. According to the accounts of school people in Dallas, neither administrators nor faculty understood the system, thus making it irrelevant to the improvement of teaching. As John Cole, Texas Federation of Teachers president, noted,

I do not believe we will motivate employees to work harder or to improve their skills by implementing an accountability program which is so complex that no one understands it. If the administration cannot even explain how it works, how do we expect school principals to advise teachers on what teaching behaviors to change in order to improve performance? (Cole, 1995)

Other important concerns included the mismatch between Dallas's curriculum and the test measures used, the absence of measures for many subject areas, and the fact that the existing measures bear little relation to the kinds of higher order learning most policymakers and educators now feel are essential (see Sykes's contribution to

this book [Chapter 10]). By raising the stakes for scores on these limited measures, the accountability system would only increase the negative effects of teaching to the test, especially for low-scoring schools or students (Madaus, West, Harmon, Lomax, & Viator, 1992).

Finally, despite the use of statistically complex, test-based accountability systems in Dallas for more than a decade (this one was only the latest in a string of such efforts), the schools there are widely acknowledged as poor, and they continue to perform badly both nationally and regionally. This may have to do with the fact that the district has expended more effort perfecting its indicator systems than it has improving schools or attracting, supporting, and keeping well-qualified teachers and other staff.

Unlike other Texas cities, such as San Antonio, that have focused on improving the training, recruitment, and mentoring of teachers, Dallas continually hires large numbers of untrained teachers and runs an alternative route to certification that produces less effective teachers than do university-based teacher education programs (Gomez & Grobe, 1990). Professional development is thin and typically focused on implementation of packaged programs rather than on skillful analysis of teaching and learning. Rather than develop professional competence and accountability, Dallas clings to bureaucratic methods—a top-down, "mechanical" approach to management (see Chapter 10) of which such accountability systems are a centerpiece. Among teachers in Dallas these approaches to accountability are seen not as a spur to better practice but rather as a spur for those with talent and options to leave for the suburbs as quickly as possible (J. Cole, personal communication, 1994).

Even if Dallas were to address all of the statistical concerns that might ever be raised about its regression equations, the question would remain: Why bother? Teaching and learning in Dallas would most likely be substantially better off without this kind of "accountability."

The Tennessee Value-Added Assessment System

The "why bother?" question is a bit easier to answer for the Tennessee assessment system, which already has proven itself useful for addressing some important research questions. The finding noted in Sanders, Saxton, and Horn's contribution to this book (Chapter 13)

that students lose ground on achievement when they move into the lowest grade in a new school is made possible by the longitudinal measures that the system uses to track student learning. These measures are a major strength of the system, along with the use of analytic tools that allow greater use of existing data. As a research tool, the Tennessee Value-Added Assessment System (TVAAS) has clear value.

To their credit, Sanders and colleagues note that "schools, school systems, and teachers cannot be assessed solely on the basis of TVAAS" (Chapter 13, p. 139). I would underscore their warranted caution. There are many reasons why TVAAS is inappropriate for evaluating teachers, and there are even more reasons why it is an unsound basis for the kind of incentive scheme Walberg and Paik propose (Chapter 15).

First, the realities of schools, classrooms, and communities—along with the limits of the TVAAS measures—make it inappropriate to attribute measured gains in learning to the skills and competence of specific teachers. As Darlington notes in his response (Chapter 14), students are not assigned randomly to teachers; some teachers routinely get more difficult students, either because they can "handle" them or because they rank low in the school pecking order. Furthermore, students in many U.S. schools are routinely tracked, as are their teachers (Darling-Hammond, 1995; Oakes, 1985). It is, of course, easier for those teachers who are assigned "high-track" students to get learning gains from students who teach themselves than it is to do so with students who need high levels of teaching skill and personal attention.

Home factors and personal learning needs or difficulties also make a difference in student performance independent of teaching skill. The large numbers of students who experience divorces, homelessness, violent crimes, hunger, or the deaths of relatives are, of course, affected by these experiences. Children who experience many of these things are not randomly distributed across classrooms or schools, as anyone who gets close to the real world of schooling knows. Even where troubled or educationally needy children are more rare, their presence in a class can have a major effect on how the energies of the teacher are deployed and on how other children respond. If TVAAS were to be used as the basis for personnel evaluations or, heaven help us, for rewards and firings, the already existing incentive in schools to avoid "problem" children would be exac-

erbated. More teachers would try—as many already do—to offload such children on their colleagues with less clout: new teachers, those who are less skilled, and those who have less political power. Children with the greatest needs would become educational lepers, and their needs would be even less well served than they are today.

For large-scale research, the methods used by TVAAS to handle missing data and to partial out the presumed effects of different teachers are unobjectionable. However, for personnel evaluations, such methods would be indefensible. No person should be evaluated for high-stakes decisions based on statistical assumptions rather than on actual information. In addition to the obvious problems of substituting mean scores for missing data, especially for high-mobility or low-attendance schools and classrooms, is the problem of attributing gains to specific teachers. The assumption of the TVAAS system is that, over time, the influence of previous teachers on a current score can be expected to decrease. This is not necessarily true. Particularly in the early elementary grades, the efforts of teachers who emphasize conceptual teaching and the building of higher order thinking skills often are not captured by traditional standardized tests until several years later, when more conceptual material and higher level performances are measured.

The limitations of the tests used constitute another obvious shortcoming of the TVAAS system as a tool for teacher evaluation. In addition to all of the concerns I raised earlier about what current standardized tests measure—and do not measure—they have low ceilings, making assessment of gains for high-achieving students especially problematic. Use of such measures for research does not pose the same problems that such use would pose for personnel evaluation tied to sanctions. Research already has well documented the incentives that would be created for teaching to the test and neglecting more complex work—research, projects, extended writing, experimentation—from the attachment of stakes to such measures. Not only would teachers who teach toward these important goals not be acknowledged, they would be discouraged from continuing to focus on these activities at the expense of drilling for the test (Darling-Hammond, 1991; Madaus et al., 1992; Shepard, 1991).

Finally, to whatever extent the tests measure useful learning, there always will be attribution problems. In some schools, students' gains in writing are actually produced by the social studies teacher who assigns regular research papers, and students' gains in mathe-

matics are the product of the science teacher's efforts because he or she uses mathematics in contexts that encourage deeper understanding—or vice versa. Without looking at practice, inferences about what teachers are actually doing simply cannot be made from statistical methods that make assumptions that are functional for highly aggregated analyses and unwarranted for drawing individual conclusions.

Some years ago, Susan Stodolsky (1984) wrote a piece titled "Teacher Evaluation: The Limits of Looking," which described why observations of teachers alone were inadequate for the purposes of teacher evaluation. The subtitle for the TVAAS system might well be "The Limits of Not Looking" to acknowledge the shortcomings of black box approaches to the evaluation of teaching.

The Oregon Work Sample Methodology

The Oregon work sampling approach has this to commend it: It actually looks at teaching, and it does so in the context of teachers' goals, classroom contexts, and student learning, measured in ways that attempt to link learning to the educational goals being sought. In these respects, it stands head and shoulders above all of the other approaches in this book as a means for providing sound evaluations of teaching that might also be useful in helping teachers improve.

As the responses by Airasian (Chapter 4) and Stufflebeam (Chapter 5) note, there are areas of further work to be done to make this a useful tool for formative evaluation of student teachers or veteran teachers. In my view, the most important areas for improvement concern the measures of outcomes used, the measures of practice used, and the approach to evaluating the effects of practice on outcomes.

In terms of measuring student learning, the Teacher Work Sample Methodology (TWSM) is dependent on the quality of assessments teachers can devise, which could be highly variable and fail to evaluate important kinds of learning well. As Stufflebeam notes, "It matters little whether ratings of teacher performance are associated with pupil gains in work sample performance if the teachers employ biased outcome measures or assess low-level knowledge and skills" (Chapter 5, p. 57). Furthermore, TWSM still envisions very limited kinds of performances that can be tested with a batch of items to

produce a "percentage correct" score rather than more complex performances that support higher order learning, such as designs for experiments, mathematical models, research papers, extended essays, oral presentations, or debates. In theory and in practice, these kinds of measures can be scored and improvements measured over time, but the methods of evaluation must extend beyond percentage correct indexes if they are to be used.

TWSM's measures of practice are not well described in the chapter, and they may be better developed than this brief chapter suggests. However, the self-referential quality of the TWSM system means that the teaching observed will only be that associated with the objectives the teacher has selected rather than a sample of performance assessed against some professional standard of practice. As Airasian notes, "Although it is important to align objectives, instruction, and assessment to one another, it also is important that the objectives themselves be meaningful and worthwhile" (Chapter 4, p. 47). This is equally true of the nature and quality of practice evaluated. My own experience with the evaluation of teachers for purposes of licensing and certification also leads me to wonder about the extent to which observers are applying common criteria for assessing appropriate practice, especially because the chapter makes no reference to common standards, measures, or training. Finally, there is presently no means within TWSM to link teaching practices to student outcomes. This limits the learning that can emerge from the system for teacher educators and teachers alike.

These critiques made, there is value in an approach to teacher assessment that points practitioners to the careful evaluation of practices, contexts, and outcomes, including the systematic consideration of student and teacher work. I have no doubt that this encourages teachers to reflect on their work in ways that are extremely productive for developing diagnostic habits of thinking as well as specific practices.

In this respect, TWSM resembles the new approaches to certification and licensure testing developed by the National Board for Professional Teaching Standards (NBPTS) and the Interstate New Teachers Assessment and Support Consortium (INTASC). These assessments also use work sampling methods to examine teachers' plans, instructional practices, assessment of student work, and feedback in the context of student work samples for several students whose progress is followed over time. The assessments already are

in use for decisions about advanced certification (NBPTS) and initial licensing (INTASC methods are used in Connecticut). However, they differ from TWSM in some key respects that make them more appropriate for these high-stakes decisions.

First, the assessment tasks are substantially standardized: English teachers must conduct literary discussions and submit videotapes of the lessons and their analyses of them, and they must show how they teach writing by presenting several assignments with student work samples and feedback over time. Middle grade teachers must outline and conduct several-week interdisciplinary units that tie several fields of study together. Mathematics teachers must show how they teach small group lessons using technology or manipulatives and large group lessons that focus on mathematical discourse. Second, as these descriptions suggest, the tasks and their evaluation are pegged to clear standards of practice that were developed by expert teachers, linked to emerging student standards, and validated in the field. Third, the examinations include both on-demand performance tasks (e.g., critique of a textbook or piece of educational software against criteria indicated in a task) and samples of work that derive more directly from the teachers' own work. Like other certification decisions, it is important to have some tasks preset so that evaluators can determine whether candidates have certain skills commonly accepted as essential for success in the field.

Fourth, the scoring systems are highly developed and have been validated and tested for reliability. They evaluate important dimensions of teaching and are administered by trained assessors who must achieve reliable levels of scoring ability. Fifth, a careful process of standard setting has been undertaken and validated. Finally, the examinations and scoring systems have been thoroughly vetted for evidence of bias, a critical concern for high-stakes testing, especially when the identity of candidates is known to evaluators. Whereas all of these features are important, perhaps most important is the fact that teachers who have undergone the portfolio and performance assessment processes of the NBPTS and INTASC assessments, or who have served as assessors, consistently cite their experience as the most powerful professional development of their careers. They believe they are better teachers as a result of this work. This claim cannot be made of the systems in Kentucky, Dallas, or Tennessee.

TWSM is not yet close to meeting these standards for high-stakes use as a method of licensure or merit pay. However, that does not

make it unuseful. As it evolves, it could be a valuable tool for preparing and assessing both beginning and veteran teachers. I suspect that teachers find it helpful in inquiring into their teaching. We should be mindful that many tools are *more* useful when they are not tied to stakes than they become when they have overreached their appropriate uses. Indeed, growing evidence suggests that policymakers' overreliance on high stakes often undermines reform rather than supporting improvements in practice. All of the assessment systems reviewed in this book would be used more productively in a low-stakes environment than for the high-stakes purposes that have been proposed for them.

The Seduction of Statistics and Stakes

Americans have a love affair with "hard" data and, increasingly, with tough-nosed approaches to school accountability. Policymakers and researchers who spend little time in schools and have even less knowledge of the realities of teaching and learning are wont to crank up numbers that appear to justify "taking names and kicking butt," as one eloquent analyst put it recently. A growing number of states and districts have gambled that education will improve if consequences for students and schools are attached to achievement measures.

More than a decade's worth of evidence shows, however, that simply setting test score goals and attaching sanctions to them does not result in greater learning—and sometimes produces destructive side effects. In many states and school districts, test-based sanctions have created incentives for schools to keep out or push out the most educationally needy students; large numbers of students have been retained in grade so that their scores look better, placed in special education so that their scores do not count, denied admission, or pushed out of schools to keep average scores up (Allington and McGill-Franzen, 1992; Darling-Hammond, 1991; Orfield & Ashkinaze, 1991; Smith, 1986). These strategies lead to lower levels of student learning and higher dropout rates in the long run, even though test scores appear to improve (Holmes & Matthews, 1984; Shepard & Smith, 1986). Similar outcomes have occurred in medicine. In one state that decided to rank cardiac surgeons based on the outcomes of their surgeries, many doctors began to turn away the most difficult

cases or to refer them to physicians in other states because they would detract from the outcome measures.

These results are predictable when accountability schemes assume that all clients are alike or that punishments are the best way to produce change. As policy analyst Bill Clune has noted, "beating the dog harder" does not lead to meaningful reform. If educators do not have the knowledge or resources to achieve higher outcomes, simply telling them to do so cannot work. In New York, for example, test-based accountability reforms of the early 1980s have caused the state's graduation rates to drop to 45th in the nation while overall achievement has declined because mandates for testing were not coupled with investments in improved teaching. If we really want accountability that works for students—rather than primarily for politicians—then we need to attend to the conditions that would actually produce better schooling and more responsive teaching.

Solving the Real Problems of Underachievement

For those closely engaged in work with schools, it is clear that the failures of recent reforms are due not to schools' unwillingness to improve but rather to the fact that most educators do not know how to teach so that all students can learn and that the systems within which they work are not structured to help them do so. If today's focus on higher standards is to result in greater student learning rather than greater levels of failure, it will need to be coupled with accountability policies that ensure that teachers and other educators have the knowledge and skills they need to teach effectively and that help schools evaluate and reshape their practices.

Achieving these goals will require approaches to assessment and accountability that incorporate at least the following.

- *Methods for examining the work of teaching and schooling*, including the work of students, using standards of practice that can propel learning and change. These include strategies such as the Oregon TWSM and assessments such as those of the NBPTS and INTASC. They also include school assessment strategies such as the school quality reviews that practitioners have found to be helpful in California, Illinois, and New York

(Ancess, 1996) as well as opportunities for teachers to collectively evaluate student work and their own practice (Darling-Hammond & McLaughlin, 1995; Murnane & Levy, 1996).

- *Measures of student learning* that are sufficiently sophisticated to assess higher order thinking and performance skills and that can allow for longitudinal evaluations of individual student progress rather than aggregated cross-sectional school scores that reflect changes in student populations more than gains in learning or improvements in teaching.
- *Means for linking assessments of learning to assessments of teaching and schooling* in meaningful ways. This will require a merger of qualitative and quantitative measures of schooling processes and "outcomes" that is grounded in an understanding of professional standards of practice as well as appropriate statistical techniques.

Perhaps in the next round of work on school accountability, we can develop tools that support student learning and the development of quality teaching as they clarify what students and teachers are doing to achieve their goals.

References

- Allington, R. L., & McGill-Franzen, A. (1992). Unintended effects of educational reform in New York. *Educational Policy*, 6, 397-414.
- Ancess, J. (1996). *Outside/inside, inside/outside: Developing and implementing the school quality review*. New York: Columbia University, Teachers College, National Center for Restructuring Education, School, and Teaching.
- Appalachia Educational Laboratory. (1996). Five years of reform in rural Kentucky. In *Notes from the field: Educational reform in rural Kentucky* (Vol. 5). Charleston, WV: Author.
- Cole, J. (1995, April 17). Letter to Sandy Kress, president, Board of Trustees, Dallas Independent School District. Used with permission.
- Darling-Hammond, L. (1991). The implications of testing policy for quality and equality. *Phi Delta Kappan*, 17, 220-225.
- Darling-Hammond, L. (1995). Inequality and access to knowledge. In J. Banks (Ed.), *The handbook of research on multicultural education*. New York: Macmillan.

- Darling-Hammond, L., & Falk, B. (in press). Policy for authentic assessment. In A. L. Goodwin (Ed.), *Assessment for equity and inclusion: Embracing all our children*. New York: Routledge.
- Darling-Hammond, L., & McLaughlin, M. W. (1995). Policies that support professional development in an era of reform. *Phi Delta Kappan*, 76, 597-604.
- Gomez, D. L., & Grobe, R. P. (1990, April). *Three years of alternative certification in Dallas: Where are we?* Paper presented at the annual meeting of the American Educational Research Association, Boston.
- Hambleton, R., Jaeger, R., Koretz, D., Linn, R., Millman, J., & Phillips, S. (1995). *Review of the measurement quality of the Kentucky Instructional Results Information System 1991-1994*. Frankfort, KY: Office of Educational Accountability.
- Holmes, C. T., & Matthews, K. M. (1984). The effects of nonpromotion on elementary and junior high school pupils: A meta-analysis. *Review of Educational Research*, 54, 225-236.
- Koretz, D., Stetcher, B., & Deibert, D. (1992). *The Vermont Assessment Program: Interim report on implementation and impact, 1991-92 school year* (CSE Technical Report 350). Los Angeles: University of California, Los Angeles, Center for Research on Evaluation, Standards, and Testing.
- Madaus, G., West, M., Harmon, M., Lomax, R., & Viator, K. (1992). *The influence of testing on teaching math and science in Grades 4-12*. Boston: Boston College, Center for the Study of Testing, Evaluation, and Educational Policy.
- Murman, R., & Levy, F. (1996). *Teaching the new basic skills*. New York: Free Press.
- Oakes, J. (1985). *Keeping track: How schools structure inequality*. New Haven, CT: Yale University Press.
- Orfield, G., & Ashkinaze, C. (1991). *The closing door: Conservative policy and black opportunity*. Chicago: University of Chicago Press.
- Pechone, R., Falk, B., & Darling-Hammond, L. (in press). *Technical report on the 1996 pilot assessments in New York State*. New York: Columbia University, Teachers College, National Center for Restructuring Education, Schools, and Teaching.
- Resnick, L. (1987). *Education and learning to think*. Washington, DC: National Academy Press.
- Shea, J. H. (1993). Deceiving oneself and others with multiple-choice exams. *The Physics Teacher*, 31, 8.
- Shepard, L. (1991). Will national tests improve learning? *Phi Delta Kappan*, 73, 232-238.

- Shepard, L., & Smith, M. L. (1986, November). Synthesis of research on school readiness and kindergarten retention. *Educational Leadership*, pp. 78-86.
- Smith, F. (1986). *High school admission and the improvement of schooling*. New York: New York City Board of Education.
- Stodolsky, S. (1984). Teacher evaluation: The limits of looking. *Educational Researcher*, 13(4), 11-18.
- Whitford, B. L., & Jones, K. (in press). Assessment and accountability in Kentucky: How high stakes affect teaching and learning. In A. Hargreaves, A. Lieberman, M. Fullan, & D. Hopkins (Eds.), *International handbook of educational change*. Netherlands: Kluwer.

BBC ONLINE NETWORK

HOMEPAGE | SITEMAP | SCHEDULES | BBC INFORMATION | BBC EDUCATION | BBC WORLD SERVICE

BBC NEWS

News in Audio News in Video Newyddion Hovosti Noticias 国际新闻 国际新闻

[Front Page](#)[World](#)[UK](#)[UK Politics](#)[Business](#)[Sci/Tech](#)[Health](#)[Education](#)[Sport](#)[Entertainment](#)[Talking Point](#)[In Depth](#)[On Air](#)[Archive](#)[Feedback](#)[Low Graphics](#)[Help](#)

Thursday, September 23, 1999 Published at 12:14 GMT 13:14 UK

How performance would be assessed



Teachers would have to get used to being monitored

The government has told the teachers' pay review body that it has been consulting teachers and others informally on how it proposes that the annual appraisals for all teachers would work.

The proposals are in a draft Performance Management Framework which the education department says will shortly be sent to all schools in England, with a similar document going out in Wales.

The draft document says: "Performance management works best when it is an integral part of a school's culture" and "is seen to be fair and open".

It also says pupils will benefit because there is "strong evidence that where schools and individual teachers are clear about what they expect pupils to achieve, standards rise".

At present, teachers progress up what is known as the pay 'spine' by annual increments to point nine on the scale - currently worth £23,193.

Changes

The education secretary envisages that things would stay much the same under the new system - unless the annual reviews of performance "indicate otherwise".

Teachers showing exceptional performance should be able to step up two pay points.

If performance is poor, the increment could be withheld and the teacher should get training and support to improve.

What teachers might have to achieve

The document gives examples of "challenging but

Relevant Stories

23 Sep 99 | Green Paper
[Performance pay: Head to head](#)

23 Sep 99 | Green Paper
[Teachers to be paid by results](#)

In this section

[Teachers to refuse 10% rise](#)

[Merit pay for best teachers](#)

[Mixed response to pay proposals](#)

[How performance would be assessed](#)

[Performance pay: Head to head](#)

[Peace breaks out in performance pay dispute](#)

[Warning over teachers' performance pay](#)

[Performance pay moves a step nearer](#)

[How teachers can earn £35,000+](#)

[Heads snub Blair over merit pay](#)

[Teachers vote to strike if pay talks fail](#)

[Pay promise for deputy heads](#)

[Teachers vote on industrial action](#)

[Union backs dialogue over teachers' pay](#)

[Concession on pay reforms](#)

[Pay: What the fuss is about](#)

[Teachers reject £1bn pay package](#)

[Government on collision course with teachers](#)

[Teachers back industrial action](#)

[Pay plan 'will not be imposed'](#)

[Warning over teachers'](#)

The document gives examples of "challenging but realistic" objectives which teachers could be set in annual assessments of their performance:

1. Objective relating to pupils' performance for a class teacher

Tests show 60% of children, and 70% of boys, in a class of nine-year-olds, cannot do what the government's literacy strategy expects.

By the end of the following year, the teacher should substantially increase the proportion of children matching expectations, and reduce the gender gap.

2. Objective relating to pupils' performance for a head of maths in a secondary school

For two years, the numbers of children receiving top GCSE grades have not reflected their achievement in national curriculum tests two years earlier.

By the year 2000, the proportion of A* and A-grades at the school should increase to 15% of entries, from 8% in 1998.

3. Objective relating to pupils' performance for a primary school science co-ordinator

The school development plan has identified improving science teaching as a priority. Attainment in physical sciences is lower than biological sciences and several staff have expressed a lack of confidence in teaching science.

By 2000, attainment in biological sciences should match that in physical sciences.

4. Objective relating to pupils' performance for a special school deputy head

Ofsted has identified a need to improve pupil behaviour.

The deputy head should draw up a plan to increase, within a year, the proportion of classes where behaviour is acceptable.

5. Objective relating to pupils' performance for a class teacher at a special school

In 1998, a class of six pupils aged seven to nine were assessed for their ability to recognise letters. One third could recognise some.

By the end of the next academic year, five out of the six pupils should be able to recognise all lower case letters.

6. Professional development objective for a primary English co-ordinator

An experienced class teacher and English co-ordinator wants to develop skills ahead of applying for senior management posts.

He or she would be expected over the coming year to develop leadership and management skills.

7. Objective relating to social inclusion for a

Warning over teachers' pay

Performance pay 'unmanageable'

Performance pay 'will not mean red tape'

Heads say performance pay is unrealistic

Ballot demanded for teachers' performance pay

Teachers 'reject payment by results'

Performance pay 'doomed to fail'

Teachers 'on the way to higher pay'

Teachers threaten appraisal boycott

secondary school year head

Truancy increases sharply at a school among 13-year-olds, who fail to fulfil potential identified by their primary school.

Within a year, the attendance of 13-year-olds at the school should improve from 85% to 90%.

8. Objective for secondary school head or deputy

At a multi-ethnic school, some ethnic groups are under-achieving.

Over a year, the teacher would be judged on whether the attainment of under-achieving groups was raised and whether exclusions were reduced.

[Advanced options](#) | [Search tips](#)[Back to top](#) | [BBC News Home](#) | [BBC Homepage](#) | ©