DOCUMENT RESUME

ED 363 619 TM 020 462

AUTHOR Wolf, Dennie Palmer; Reardon, Sean F.
TITLE Equity in the Design of Performance Assessments: A Handle To Wind up the Tongue with?
SPONS AGENCY Rockefeller Foundation, New York, N.Y.
PUB DATE Mar 93
NOTE 62p.; Paper presented at the Ford Foundation National Symposium on Equity and Educational Testing and Assessment (Washington, DC, March 11-12, 1993).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Academic Standards; Access to Education; *Educational Assessment; Educational Attitudes; Educational History; Educational Policy; Elementary Secondary Education; Equal Education; *Evaluation Methods; Excellence in Education; Intermediate Grades; Junior High Schools; *Material Development; Middle Schools; National Programs; School Effectiveness; State Programs; Urban Schools
IDENTIFIERS Opportunity to Learn; *Performance Based Evaluation; *Project Performance Assess Collaborative for Educ

ABSTRACT

In discussing educational equity and Project Performance and Assessment Collaboratives for Education (PACE), the authors present an argument, a detailed illustration, and a set of imperatives for education in the United States. The argument examines the last century and a half of American educational practice to illustrate why we cannot simply assume that setting high standards and creating performance assessments will automatically move us toward educational equity. Our national problematic view of intelligence as something that never changes and our difficulty in defining what excellence really means require that we establish new ways of talking about learning. The argument is illustrated by examining the work of urban middle schools that are members of the PACE Project, funded by the Rockefeller Foundation to expand access to opportunities to learn. The role that a coherent system of performance assessments might play in reducing differentials in this population is discussed, focusing on curriculum-embedded assessments as powerful tools for modeling, enhancing, and yielding evidence about opportunities. Implications for educational policy at national and state levels are discussed. However useful regulations to promote equity are, they are no substitute for the realization of equity in day-to-day decisions and interactions in the schools. Three figures provide examples from assessment practice. (Contains 58 references.) (SLD)

Equity in the Design of Performance Assessments:

A Handle to Wind up the Tongue With?

Dennie Palmer Wolf and Sean F. Reardon

Paper prepared for presentation at the Ford Foundation
Symposium on Equity and Educational Testing and Assessment,
Washington, DC, March 11-12, 1993

The writing of this paper was supported with a grant from the School Reform Project of the Rockefeller Foundation which has funded us to work with a national network of middle schools. In particular, we would like to thank Linda Carstens, Edmund Gordon, and the teachers and students in those schools for helping us to understand how deliberate, rather than assumed, our approach to equity must be.

OVERVIEW

Only three quarters of a year separate the announcement of President Bush's "America 2000" proposal in April, 1991 and the January, 1992 National Council on Education Testing and Standards' report to Congress which concludes that a national testing system is both "desirable and feasible" (National Council on Education Standards and Testing, 1992). But this endorsement is less a reasoned conclusion than a triple wager: First, it is a gamble that if, as never before, we link assessment to high standards, we can galvanize significant and lasting change in American public education. Second, it is a bet that the effect of setting high common standards will, in and of itself, be constitutive of equity. Finally, at least several of the ensuing plans for such a national testing system (NCEST, 1992; New Standards, 1992) carry the additional hope that via the criterion-referenced performance assessments such a system will force wider access to worthwhile educational activities like writing, problem-solving, and scientific experimentation.

But in all the heat and light of this debate, and a flurry of technical questions, a fundamental question almost escaped: Neither the profoundly cultural act of standard-setting, nor a technology like testing, is a neutral tool. Exactly like printing or nuclear power, the value of these new assessment technologies depends wholly two prior matters: plausibility and use. How plausible is it that we can create anything but a false average of high standards in a world in a world of savage inequalities (Kozol, 1991) where a student in urban Hartford cannot access what her suburban peers have by right of address? In addition, we have to ask: "**Used in what ways** could standards and performance assessments promote more equitable access to educational opportunities?"

In this paper, we present an argument, a detailed illustration, and what we see as a set of imperatives. The argument examines the last century and a half of American educational practice to illustrate why it is we cannot simply assume that the acts of setting high standards and creating attendant performance assessments will, in and of themselves, move us closer to educational equity. In making this argument we examine two chief sources of concern: our particularly vexed views of intelligence and excellence, as well as our fluctuating attention to issues of educational equity. In our illustration, we turn to the work of a set of urban and rural middle schools where the concern for equity is, for many reasons, sharp. These schools are members of the PACE Project, a national network of schools, funded by the Rockefeller foundation to work on expanding and diversifying the population that has access to "opportunities to learn" - to think, imagine, and invent (Smith & O'Day, in press). In this illustration, we look, in detail, at the role that a coherent system of performance assessments might - if it were so designed - play in countermanding differential (Smith & O'Day, in press; Wolf & Baron, in press). In particular, we examine the uses of curriculum-embedded assessments as powerful tools for modeling, enhancing, and yielding evidence about these opportunities to learn for all students.

In closing, we discuss the implications for policy at the national and state level. For instance, no participation in a national, or regional, or state assessment program, that is not accompanied by a parallel assessment of opportunities to learn and a school-based reading of performance data in light of equity concerns. However, our most emphatic point is that however comforting such a "corset" of regulations may be, it is no substitute for the realization of equity in the moment-to-moment interactions, and day-to-day decisions of schools. Therefore, we argue that it is crucial for government, foundations, and communities to "think small" and in so doing

to support the sustained development of actual public schools where many children have opportunities to learn and where what they learn is wisely and productively assessed. At the same time, we argue for the creation of a networks of *horizontal* accountability, in which schools participate in networks of peer-review.

In her novel, Their eyes were watching God, Zora Neale Hurston reflects on language we come to use all too easily. She remarks on how the town of Eatonville comes to speak of their mayor and her husband, Joe Sparks:

> There was no doubt that the town respected him and even admired him in a way. But any man who walks in the way of power and property is bound to meet hate. So when speakers stood up when the occasion demanded and said "Our beloved Mayor," it was one of those statements that everybody says but nobody actually believes, like "God is everywhere." It was just a handle to wind up the tongue with. (Hurston, 1937, p.77)

We have, all of us, learned to say - and to write - "equality and excellence," or "high, common standards for all students." Such phrases have become virtual passports into discussions of school reform. The real question is whether or not we are willing to do the work and enter the debates that will turn "handle to wind up the tongue with" into messy, flawed, but actual practices.

THE TROUBLED LEGACY: INTELLECT, EXCELLENCE, AND EQUITY IN AMERICAN PUBLIC EDUCATION

Problematic Views of Intelligence

We measure, test, and evaluate according to our socially constructed ideas about what constitutes knowledge, learning, and intellect. Consequently, the fairness of assessment practices cannot be meaningfully divorced from the epistemology that underlies them; the castle is only as solid as the cloud on which it rests.[1]

Intelligence as Fixed and Unevenly Distributed

In the United States, the history of educational assessment is full of instances of the use of tests to limit access to opportunity for women, immigrants, and Blacks. Beginning with the development of the first group intelligence tests during World War I, group differences in average test scores were used to justify denying educational opportunities to Blacks and limiting immigration from Eastern and Southern European countries (Gould, 1981; Kamin, 1974). Well into the twentieth century, segregationists used test score data to argue that Blacks and Whites could not be educated together (see, e.g. *Stell* v. *Savannah-Chatham*, 1963). It is important to realize, in addition, that discriminatory actions based on the results of intelligence tests held legitimacy not merely because the test results reinforced dominant social attitudes, but because they gave those attitudes a patina of "scientific" -- and hence, unassailable -- validity (Graham,

[1]Much of the following section is taken from the discussion of these issues in Wolf, Bixby, Glenn, & Gardner (1991), pp. 36-47.

1993).

Underlying most testing practices in the U.S. is an epistemology based on three "highly debatable" assumptions (Wolf, Bixby, Glenn, & Gardner, 1991, p. 37). Early intelligence tests were assumed to measure a unitary, quantifiable trait called "innate intelligence" or aptitude, as it later came to be called, rather than criterion-referenced achievement. This innate intelligence was assumed to be heritable and immutable; and correlations between race, class, and gender and scores on intelligence tests were taken as evidence of between-group genetic differences in naturally occurring-intelligence. Thus, as Terman so confidently described, individuals could - and should - be ranked in stable ways according to their mental capacities and education should be tailored to these differences so as to prepare individuals for their inevitable lot in later life:

> Preliminary investigations indicate that an I.Q. below 70 rarely permits anything better than unskilled labor; that the range of 70-80 is pre-eminently that of semi-skilled labor; from 80-100 that of the skilled or ordinary clerical labor; from 100-110 or 115 that of the semi-professional pursuits; and that above all these are the grades of intelligence which permit one to enter the professions or the large fields of business. Intelligence tests can tell us whether a child's native brightness corresponds more nearly to the median (or one or another of these classes). This information will be of great value in planning the education of a particular child and also in planning the differentiated curriculum. (Terman, 1922)

The resulting confusion of achievement with ability has meant that social groups with less access to school-based knowledge have been historically viewed as intellectually deficient. This, in turn, has led to patterns of differential access to curriculum, instruction, expectations, and even school

rewards that, in practice, insure and reproduce differential levels of achievement (Fass, 1989; Oakes, 1985; Wolf, Bixby, Glenn, & Gardner, 1991).

Constrained Views of Excellence

The curriculum and goals of American public schools were also heir to a particular notion of human change over time -- that of linear and uni-dimensional progress. This is hardly accidental: medieval and Renaissance accounts of society are based on a great chain of being leading from serf to monarch (Gould, 1981; Kamin, 1974). Shakespeare's history plays portray an orderly and productive world as one in which just such chain of status prevails. Authors like Dante and Milton -- fundamental architects of our imagery and language -- envision the good life as progressing from the lower depths of Hell, struggling through Purgatory, and ascending finally to Paradise. This view of human social order became a model for a highly linear epistemology in the early seventeenth century with Descartes, Leibnitz and others who argued that access to a set of general and abstract rules, superseding particulars, were the most powerful route to religious ecumenism, scientific progress and intellectual power (Toulmin, 1990). The notion of single, rather than multiple and noisy pathways became social policy in the nineteenth century when it informed (and justified) colonialism. It is a view that infused G. Stanley Hall's argument that children evolve from natives to citizens. (It still permeates the field of child development through the concept of a fixed sequence of stages of mental, moral, and emotional development, many of which insist that maturity or adulthood can be defined as operating at the level of rational, general rules, rather than with an embeddedness in the "noise" of particulars.)

This view of development has a complement in a stiffly hierarchical view of kinds of

knowledge. As far back as Aristotle, philosophers have drawn a sharp distinction between the 'superior' work of thinkers who ask why and the 'inferior' work of artisans who make and do. Subsequent theories of knowledge, from Descartes to modern quantum physics, have continued to privilege acts of pure thought. In this view, theory-building, the acquisition of concepts, general rules, and symbolic manipulations are more worthy than practical, situated or commonplace problem-solving (Resnick, 1987; Scribner, 1984; Toulmin, 1990). Translated into school terms, students who can say "he runs" or "I run" don't understand subject-verb agreement. Only those who can state the rule and define the terms "really" grasp the concept. Thus, there is a distinct hierarchy of knowledges: at the bottom is practical problem-solving embedded in the particulars of actual human situations, at the upper reaches is theoretical speculation and knowledge of the general case (Toulmin, 1990). One result is that we have selection of *a* very particular set of displays as evidence of having reached those valued outcomes of rational, symbolic and abstract thought -- algorithmic mathematics leading to calculus, essayist prose, reading and translating rather than speaking a foreign language, laboratory science over field studies, statistics and models over narrative, and representational drawing over design. Via testing, these modes of display have become the definition, even the reification of the skills for which they stand. This has led us to the belief that if a student cannot write a plot summary or a "compare and contrast essay," he has not understood the particular reading passages or documents or novel in question. But, in so doing, we closed the doors on any substantial discussion of the fact that those displays were ways to tap understanding, not the understanding itself.

The Consequences of these Views of Intellect and Excellence

These points are not just academic subtleties. Like our epistemology, our notions of excellence have deep-running, and very practical consequences. The first has to do with our considerable difficulties in distinguishing difference from deficit. In effect, we installed an emphatically narrow vision of excellence at just the time when those institutions opened their doors to girls, to African-American families travelling to Northern cities, and to immigrant families (Cuban, 1984). The consequence is a situation in which teachers can distinguish between ranks of performance - so long as those performances come out of the same conventional and cultural cloth. Consider these in-class responses to a story, "Jemo Shinda," in which a young boy is run over by the careless drivers of an open car, reminding the author of the wild, cruel abandon in The great Gatsby.

> Its about a boy gettin run over by a car. It runs him over and his sister whos the writing the story tries to get there father and mother but he dies anyway. They were Japanes living in America for the first time so it makes her think she didn't like living here so much. Even if it is America.
>
> "Jemo Shinda" is a short story by Hisaye Yammamoto. In this story the author tells about the time when her brother was run over by a fast car. By doing that, she makes you see the roaring 20's were not fun and games for everybody.
> (Wolf, 1992, p. 6)

We can give the first a C, based on its plainess, its grammar, and its sticking to little more than partial summary. We can give the second response a B+, because it is better on all those same and familiar dimensions. But contemporary classrooms contain a much wider range of responses. Some are not easy to place on the familiar scale. For instance, this response written by a student who, in high school, is still learning English.

> The author, her name is Hisaye Yammato (sic) tell how the rich shiney (sic) cars coming home full of partying and drinking, they ran over a boy so poor and so foreign he was no different from their flat driverways or soft grass. It makes me to know when Mr. Fitzgerald write about the roaring 20's, it is roaring like a hungry animal, not like singing.

A piece such as this, is problematic. It is striking and the conventions are poor. In truth, it is hard to tell whether the imagery is intended or the result of hunting for ways to use a limited vocabulary and syntax to convey what, at a later point in English acquisition, the student may sum up simply with a word like "savage."

The second cost is the loss of diagnostic power. Insofar as we view individuals (or their performances) as being at *a* stage, we fail to communicate the message that any human learner or any complex performance is composite. We cannot pick out for an individual the pattern of her strengths and weaknesses with the result that educational conversation becomes a blunt and clumsy tool. A case in point is the 1990 NAEP assessment of classroom narrative writing . That study ranks all performances on *a* single scale that runs from "event description" to "elaborated

story." The scale, in effect, presumes a single, and a parochial, endpoint: Any story worth its salt should take the form "we" know from Aesop or Grimms: a hero with a problem to solve who solves it. To the extent that a performance veers from that expected norm, it is deviant, or "unscorable." (Thus, the NAEP scoring system awards only a four a compact account of what happens to a rider when he looses control of his new bike on a downhill incline. By comparison, a student's thorough retelling of a Halloween movie plot, receives a score of six. Gabriel Garcia Marquez or Yasunari Kawabata, given their fondness for ellipsis and ambiguity could do quite poorly on this scale. In addition, the independent dimensions of quality of language, the originality of plot, the vividness of characters, the capacity to borrow from reading or to include details borrowed from personal experience are all submerged (National Center for Education Statistics, 1990). There is no question about the efficiency, and perhaps even of the psychometric rationale behind such an approach. However, in promulgating it, we throw away a remarkable opportunity to *discuss* the *several* aspects of performance that make for strong and for distinctive work. As one student remarked, after struggling to score a peer's paper on a six point holistic writing scale, asked, "But what do I do if she is not a three straight across?"

Still another cost has yet to play itself out fully. We have kept the forms of displaying understanding absolutely steady for a century, sticking for instance to the long, complex displays of calculation as evidence of mathematical skill, even as computers and calculators have made that an increasingly irrelevant proxy of interest in or skill at mathematics. We still hold out essays as the honored form of display in history classes, where increasingly much work is being done in formation of hypercard stacks. Technology is not the only source of strain: the world around schools is changing. Foreign language learning is no longer a drawing room or Grand

Tour skill. Students need it because many of the patients they care for, or the clients they serve will speak Spanish, Mandarin, or Japanese. Consequently, oral communication may matter much more than the translation of literature.

Toward a Language of Development and Varieties of Excellence

It is premature at best to worry about equity in performance assessment, without taking on prior issues: notably, the underlying belief that intelligence never develops and the still-open questions about how to recognize excellence in a world no longer simply constituted only of the young men of St. Paul's, Exeter, and Bronx High School of Science wielding fountain pens.

We have to invent what amounts to different paradigms or images for talking about student learning. Quite possibly, we should take a page out of the work of New Zealand and Australia, where they have described major benchmark achievements in domains like literacy: 10 bands stretching from pre-kindergarten to twelfth grade. Teachers have agreed on a range of bands appropriate to each grade level, and report on where students stand with respect to that criterion. In addition, however, they report the range of a student's work, across a semester or year, so as to picture development, not just snap-shot achievement. As educational units, each grade takes seriously the work of condensing the range of performance and moving the bulk of student performance to the high end of the range.

We have to re-think excellence, not as the upper tail of the distribution of performances (or as what the upper 5% have), recasting it as performance at a high standard. That should have substantial and detailed effects. We want standards that are high and that are common. But we also must acknowledge that standards can be met variously. We have to ask, for instance, if

where we have always demanded an essay, a student could write a fictional last chapter to a book. Or if by illustrating a story and arguing for the choices made in those illustrations, we might not also get at comprehension. In this light, we might also consider stealing from the arts and humanities: The point - in piano competitions or reviews - is agreement on enough common language to talk responsibly about basics (tone, getting all the notes, playing challenging-enough repertoire, etc.) but then to allow for the kind of discussion that comes from different traditions, values, and tastes. Students leave us for a world were baseline tolerances are set (the bore on a particular piece of machinery must be within .025 of a millimeter, but where large questions of national spending priorities, who wins an election, or who is guilty as charged is full of politics, style, and beliefs play into the ongoing debate about "what is good." The educational discussion of excellence, at least in publically-funded schools, has been shaped considerably by a hunt for just those categories on which it is possible to create extensive agreement. One of the consequences is that many scoring systems turn to easy countables (how long, how many items are correct, how much detail is there in the writing, how much evidence is present in the history essay). Or the scoring systems that are about covert counting (such as the NAEP instance described above.) Perhaps we ought to explore systems that are deliberately bi-focal: yes, common categories and agreement for them, but also a deliberate place for narrative comments that give a place to taste, beliefs and values, or fashion. Students leave schools for a world in which only the most basic tolerances are set, but where there is large play to matters of taste. If we want students to be able to make wise use of the interplay of responses their work is bound to evoke, don't they need to be invited into that conversation? It is the essence of peer review. And the process through which a culture refines and evolves its ideas of what is "a winner," "a

break-through," or "a dud."

Let us be crystal clear here. We are *not* proposing what many will rush in to accuse us of: a loosening and lowering of standards, a kind of lazy "let a thousand flowers bloom." Yes, to high and common standards. But in an examined way. In the long run no sustained assault on the inequities of our current system can proceed without individuals who are trained to debate, rather than to ratify comfortable answers to "what counts as evidence of understanding." We are also not naive. The decisions about what is "good" or "good enough" is closely guarded cultural capital. Institutions are built on it. Job interviews are fashioned on it. We are not proposing to create a generation of students who can only speak, not write; who can make videos but not argue. Who can estimate but not measure or calculate. But we do want to ask, out loud, "Isn't it possible to acknowledge that there are multiple, equally steep routes to understanding and at least more than a currently anointed ways of displaying of knowledge?

The Rise and Fall of Equity on the National Educational Agenda

Given our problematic views of intellect, and our stubborn views of excellence, it is not surprising that we inherit a troublesome history of equity in education. What such an examination reveals is a half century of a quite checkered, and often shallow, commitment to the idea of high common standards for all students.

Prior to the 1954 *Brown* v. *Board of Education* Supreme Court decision, the idea of educational equity in the United States, insofar as it existed, was based on the "separate-but-equal" doctrine of the 1896 *Plessy* v. *Ferguson* decision. In theory, all students had access to the same educational opportunities, albeit in separate and segregated schools. In practice,

however, students differing in race, ethnicity, and social class received very different educations, based on society's perception of their "evident or probable destinies."[2]

Issues of educational equity gained serious attention in the United States only in the years following the 1954 *Brown* v. *Board of Education* Supreme Court decision. Thurgood Marshall argued before the Supreme Court in *Brown* that separate was inherently unequal; Blacks could never have the opportunity to succeed if they were subject to a segregated educational system (Kluger, 1975). In the civil rights era, evidence of racial differences in test performance--evidence that had previously been touted as evidence of Black inferiority (see, e.g., Gould, 1981; Kamin, 1974) -- was increasingly interpreted as evidence of unequal opportunity (though there were still those who argued for a genetic cause for the differences -- see, e.g. Jensen, 1969; *Stell* v. *Savannah-Chatham*, 1963). Early legal interpretations of the *Brown* decision focused on ensuring formal equity of inputs (Liebman, 1990); the ensuing equity debates of the 1960s centered therefore on issues of equal access and on the even distribution of easily quantifiable inputs -- ready-to-hand countables such as racial balance in schools, matched facilities, equivalent levels of teacher preparation, and equal funding.

The assumption that equal opportunity would eventually lead to social and economic equality led the federal government and federal courts to focus their attention on removing obstacles to equal opportunity -- and opportunity was generally defined as access to an integrated educational environment. The many federal desegregation orders, the civil rights laws of the 1960s, and the War on Poverty programs -- Headstart, Job Corps, and federal aid programs for

---

[2]The phrase is from a 1908 speech to educators by Charles William Eliot, then-president of Harvard University (quoted in Graham, 1993, p. 90).

schools and colleges -- were all based on the understanding that racial discrimination and poverty were at the root of inequalities and that basic change required outside coercion, incentives, or the creation of new institutions whose basic function was to expand opportunity (Orfield & Reardon, 1992; 1993). Thus, the 20 years following *Brown* saw the most substantial drives toward equity that this country has known. However, even in that climate, concern focused on improving access to educational institutions for minorities and the poor, not on what Goodlad was later to call, "access to understanding" (Goodlad & Keating, 1990).

The nation's attention to educational equity issues, however, has waned considerably since the early 1970s. This decline has not been due to sharp improvements in the equality of opportunity, but rather to the shifting political environment. Despite continued evidence of obvious inequalities in education there were no important initiatives for equity under way either in government or in the courts throughout the 1970s and 1980s. Republicans dominated the presidency and controlled court appointments for all but four years from 1968-1992, effectively shifting the national agenda away from equity concerns. In response, Democratic presidential candidates adopted a strategy aimed at winning back the increasingly powerful White/suburban/middle-class vote by downplaying issues of race and equity. With neither political party willing to champion policies aimed at increasing equity and opportunity for low-income and urban residents, and with the Republicans actively engaged in dismantling previous efforts at improving equity, the concerns with educational equity, access, and opportunity that had been prominent in the civil rights and poverty struggles of the 1960s were largely forgotten (Orfield & Reardon, 1992).

Equity in the Excellence Era

In recent years, as equity issues have fallen off the national agenda, educators have had to slip it in -- almost covertly -- on the coat-tails of the escalating drive for outcomes, accountability, and eventually, the discourse about excellence. The shift to a focus on outcomes actually began with the testing-for-accountability movement, and was incorporated into the calls for excellence in education of the 1980s. Widely publicized (if misleading) evidence of declining SAT scores in the 1960s and early 1970s led to a belief that failures of the educational system were responsible for the stagnant national economy of the 1970s. Control over education was increasingly centralized through the century -- by 1980 state funds accounted for 47% of all educational spending, up from 17% in 1920, and local share had dropped from 83% to 43% (Digest of Education Statistics, 1986); in the 1960s, the ESEA gave a boost to the development of what had previously been small or non-existent state-level educational authorities -- by the 1970s and 1980s virtually all states had large state educational agencies. Finally, the development in the 1960s of the technology of criterion-referenced standardized testing (Airasian, 1979) gave states a vigorous, even a tough, means to hold the schools accountable. By the mid 1970s, the conjunction of calls for public accountability of educational outcomes, the centralization of educational control, and the availability of criterion-referenced testing technology led to the educational testing reforms of the 1970s and 1980s, including minimal competency exams for grade promotion and graduation, as well as the use of mandated periodic student testing to reward or sanction teachers, schools, and districts. Driven largely by state policy makers and business leaders, these reforms were based on a manufacturing model of education, grounded in the assumption that holding schools, teachers, and even students responsible for

student outcomes would provide schools with strong incentives to improve their educational practices.[3]

Though the excellence movement was based primarily in conservative accountability reforms and gave only lip-service to equity issues in its "excellence-for-all" rhetoric, many educators and others concerned with equity found that the focus on outcomes was their only available lever for including equity in the public education debate. This was crucial at a time when the equity concerns of the 1960s -- racial integration and equal opportunity -- were losing political support (Britell, 1980, Cohen & Haney, 1980). The excellence-for-all rhetoric appealed to conservatives for its accountability focus, and to liberals because of its focus on equal outcomes (Haney & Madaus, 1979). James Liebman, a lawyer at Columbia Law School, argued in 1990 that state minimal competency standards give civil rights lawyers a new strategy for holding states responsible for improving educational equity (Liebman, 1990).

Beyond Matched Inputs: Toward Matched Opportunities to Learn

But the results from this legal hitch-hiking have been thin: a few desegregation orders, which in the 1960s tended to focus on access and input variables, have included improvement in outcome measures -- e.g., dropout rates, achievement scores -- as conditions for termination.[4]

---

[3]This description of the roots of the testing-for-accountability movement is derived from fuller discussions of its history and rationales in Airasian (1979), Brandt (1981), Britell (1980), Cohen & Haney (1980), Farrar (1990), Greene (1980), Haney & Madaus (1979), and Pedulla & Reidy (1979).

[4]Cleveland is the most notable example of this: see *Reed* v. *Rhodes*, 1978, and its subsequent court orders.

Overall, as a basis for monitoring and ensuring equity, a focus solely on outcomes is inadequate. Such a focus treats the practice of education as a black box, and implies that we know nothing about what educational practices are more effective and equitable than others. Moreover, it opens the system up to corruptibility -- if we pay no attention to how schools achieve desired outcomes, we create strong incentives for them to implement practices which may have educationally detrimental effects even as they improve apparent educational outcomes. Poignantly, high test scores do not necessarily mean that students are learning more of what we want them to (Cannell, 1987; Madaus & Kellaghan, 1992). Most importantly, to hold schools and students responsible for meeting outcome standards without providing them the conditions necessary for them to do so is unjust:

> It is not legitimate to hold students accountable unless they have been given the opportunity to learn the material of the examination. Similarly teachers or schools cannot legitimately be held accountable for how well their students do unless they have the preparation and resources to provide the students with the opportunity to learn. (Smith & O'Day, forthcoming, p. 32)

In response to these concerns, educators have increasingly argued for more attention to the *practices* of schools. Smith and O'Day, for example, define educational equity as providing all students "the opportunity to learn well the content of the [curriculum] frameworks" (p. 33), and recommend the adoption of three sets of standards: 1) *resource* standards, 2) *practice* standards (what could also be called "opportunity-to-learn standards), and 3) *student performance*

standards (p. 37). These types of standards have become increasingly common; the Connecticut State Board of Education, for example, stated in 986 -- long before the metropolitan Hartford desegregation case currently in state court -- that its goal of:

> '[e]qual educational opportunity'....require[s] *resources* to provide each child with opportunities for developing his or her intellectual abilities and special talents,....[and is evidenced by] the participation of each student in *programs appropriate to his or her needs* and the achievement by each of the state's sub-populations (as defined by such factors as wealth, race, sex, or residence) of educational *outcomes* at least equal to that of the state's student population as a whole. (Connecticut State Board of Education, Policy Statement on Equal Educational Opportunity 1, as cited in Liebman, 1990, footnote 122, italics added)

This definition of equal educational opportunity includes the three types of standards advocated by Smith and O'Day. Similarly, Liebman (1990) argues that given the failure of strategies based on achieving equity through equal access, civil rights lawyers should concentrate on achieving "equal chances" -- or matched opportunities to learn. And, in fact, a recent analysis of the relationships between eighth graders' achievement and their "opportunities to learn" -- as defined by curricular offerings and instructional methods -- shows that schools' practices play an important role in students' achievement (Epstein & Mac Iver, 1992). Any accountability system which does not include measures of such practices is likely to miss much of what actually counts towards the achievement of more than a veneer of equity.

This begins to spell out some implications: what kinds of items might we want to include when looking at whether or not schools provide equivalent, or comparable, opportunities to learn. We know, by now, that it will yield little to look at these sorts of easy countables: numbers of books in the library; numbers of post-collegiate training among teachers; etc. Instead, we would want policy-makers to turn to such difficult-to-measure, but genuine, indicators as: students' access to large ideas and worthwhile strategies (e.g., independent reading, library research, use of large data bases, etc.) We would also want quite different things such as evidence that students have access to models of work well done; that students have the ability to revise their work in the light of more recently acquired understandings, etc. Evidence that teachers had good diagnostic skills across the full range of students they taught.

THE DEED IS IN THE DOING:

DESIGNING EQUITABLE PERFORMANCE ASSESSMENTS

As part of the growing concern about what schools are actually delivering to students, many educators argued that the use of standardized tests for accountability has actually narrowed curricula and driven instruction increasingly toward pedagogies based on memorization and basic skills rather than improving educational quality. High-stakes standardized testing policies, many argue, are highly corruptible; creating greater incentives for cheating than for actually improving instruction (see, e.g., Cannell, 1987; Madaus & Kellaghan, 1992). In response, many educators have advocated more "authentic" forms of assessment, including portfolios and performance-based assessment of many kinds, coupled to high common standards rather than the low "basic

skills competencies" emphasized by the testing movement. Such assessments, many argue, would crystalize educational goals, make them transparent, and provide continuous feedback about their achievement by various groups of students. For instance, in his 1984 book *Horace's Compromise*, Sizer argued that schools should require "exhibitions of mastery" as a condition of graduation. Sizer did not challenge the belief that schools needed to be held accountable to the public, but instead argued that

> The requirement for *exhibitions of mastery* forces both students and teachers to focus on the substance of schooling. It gives the state, the parents, prospective employers, and the adolescents themselves a real reading of what a student can do. It is the only sensible basis for accountability. (1984, p. 215)

But, in the light of our concern for equity, there is an important issue here: the call for performance assessment did not derive from concerns for fairness or equal access. It arose, instead, as an alternative to standardized testing as a way of ensuring accountability and excellence. Performance assessments, educators like Sizer have argued, will focus attention on the real business of schools, and will therefore drive the curriculum upward to ensure common high standards. Having made their way onto the national educational agenda on the coattails of the excellence, it remains an open question whether or not performance assessment can be designed to serve equity.

A Well-designed Performance Assessment System

A well-designed performance assessment *system* might act as a lens to more clearly reveal existing and ongoing inequalities and to inform policy and practice in teacher training, teaching practice, curricular design, and school organization, *but only if it were explicitly designed to do so* in all its parts. But to do so requires a sharp expansion of our most familiar purposes for collecting information about student achievement. The familiar function for assessment is to yield information about what students know and are able to do: as in the case of California Achievement tests, the Degrees of Reading Power, or more demanding performance assessments. In a system seriously concerned with equity, these sorts of measures reveal whether or not schools are turning out students whose achievement is tightly correlated with their family income, gender, first language, or ethnicity. But assessment must be more than that if equity matters to us. It must become an occasion for student learning about the achievement of and the standards for good work; and an opportunity for schools or system to learn about the consequences of their current programs.

Assessment as an Occasion for Learning: The Case of Students

In the lives of students we value, the fundamental purpose of most assessment is to teach the rules of excellence, not to sort. To make this point, we turn to an example. It is an exchange between a student in a demanding American history class and his teacher over a research paper that the student wrote on the *Plessy* vs. *Ferguson* case. The student was frankly delighted with the paper, because, as he put it:

It was the first time I ever really wrote a paper with a point of view. I didn't try

> to summarize all the background, and then the plaintiff's arguments, and then the state's arguments. I wrote almost entirely about what the plaintiff's side had to argue. See, I had this theory that they lost, not just because of prejudice, but because they were working out the arguments for civil rights for the very first time and they were having to work metaphorically -- like by arguing that a person's identity as an equal citizen was like property and not to be allowed to ride in a mixed railroad car was, therefore, to steal his property.

However, the student's teacher was not bowled over. Assigning the paper a modest B+, he commented on what he thought was the unbalanced quality of the paper.

> Your paper focusses to too great a degree on the plaintiff's arguments in Plessy. Though they have bearing on the rest of your paper, the arguments take on a place in your writing that seems inordinately large. The arguments of the defense are dealt with in a page, and the actual ruling is described in a rather cursory way. The dissenting opinion merits only a sentence, despite its sensitivity to the arguments put forth at great length in your paper by the plaintiffs. Thus, it seems that as scholarly as your approach is in this research paper, there was not enough attention on your part to creating a balanced examination which would give your readers more of an idea of the dimensions of the controversy. (Cf. Figure 1 for the entire set of comments)

The student, incensed and not about to be put off, rallied, marked up the teacher's comments with comments of his own (Cf. Figure 2). In that second set of comments, the student argued fiercely for his right, as a historian, to take a "particular cut" on the issues.

> To have given equal weight to both Plessy's and the state's arguments, in my mind, would have been ridiculous. The idea that a black man was to challenge a white man's law in the Supreme Court in the late nineteenth centruy (sic), that in itself should be a clue that the black man's arguments must be stupendous and overwhelming, if he seeks to win. Furthermore, the burden of the proof was on Plessy, not the state. The state was content to stick by its law, whereas Plessy had to produce the proof that his law was unconstitutional. Therefore, Tourgee and his associates used virtually everything they could, the l4th amendment, the damaged property argument, and the precedent Brown case, to name a few. The bottom line is that this case, in my opinion simply cannot be presented as an equally argued affair. It wasn't and that's what makes it Plessy vs. Ferguson.

Responding to the student some weeks later, the teacher agreed that a totally balanced summary of the case was not the point, so much as an understanding of the issues. At the same time that he relented with an A-, the teacher also held his ground about the student writer's responsibility for using a title, his introduction and conclusion to make clear that he had a particular point to make (Cf. Figure 3).

> You have done a superb job of defending your position. Your arguments are generally persuasive and always articulate and forthright. I feel that you have given me a much clearer idea of what the objectives of your paper were. I wish those objectives had been more clearly stated in your introduction.... I accept your arguments for placing emphasis on Plessy's suit.

The point is simple: in the lives of students we value, assessment is not merely a thermometer that we use occasionally to check up on that student's achievement. It is always (if tacitly) part of the curriculum: as in the example, it teaches the nature of success. In the last century, assessment has been among the most powerful tools we have for creating unequal access -- not so much to content -- but to standards and to the strategies for improving one's work. To "ordinary" or "average" students we provide, not critical comments and response, but information about the percent of items correct, relative ranking, offered up as a single undifferentiated score, unaccompanied by any constructive response. Moreover, these scores, or grades, arrive back too late to affect their work. They are, quite simply, terminal events.

It takes the system's temperature, but it lacks much by way of causal or explanatory evidence. At most it can serve as a signalling mechanism.

Curriculum-embedded Assessments

If we want to know not just where and for whom equity concerns are being met, we have to have additional assessments that probe for reasons: just why is it that Hispanic girls are nowhere to be found in physics or that Asian-American boys seem to have more difficulty with

writing than Asian-American girls, or that students in the downtown schools score more poorly than those who live in suburb-like neighborhoods? With increasing frequency, many people are arguing that portfolios fill this bill: witness their increasing role in state assessment programs and in projects like the New Standards Project, PACE, the increasing discussion of their potential role in Chapter 1 programs. Here we want to argue for the a third ingredient in an assessment system determined to promote equity and excellence: curriculum-embedded assessments.

What is curriculum-embedded assessment? Recent research has made it clear that the familiar approach of one-shot performance tasks, given in a burst in April, cannot predict a student's performance on a novel task, because performance can vary so significantly with content and type of task. Instead, it may take between eight and ten different samples from the domain - before we could stand behind our judgements. This is critical as states and schools move toward using portfolio-like data to create achievement scores for individual students. Realizing that no teacher would want his students to sit still for assessments for four days running in order to produce the requisite data, a number of researchers have proposed banks of performance tasks from which teachers might draw ones on energy, friction, heat to administered whenever she came to the end of the appropriate stretch of work. Hence, the name: tasks would be done at the time and in the sequence that would permit them to embedded at the appropriate moment in the ongoing flow of classroom instruction.

This is the technical origin and definition of curriculum-embedded tasks. However, such tasks have other interesting and promising features that we could - if we act - convert to a deliberate, rather than accidental, tool for equity. Such episodes of assessment are multiple and

spread throughout the year. Thus they are an ongoing conduit into the curriculum: well-structured they could actually insure that large numbers of students write, discuss, and revise; gather data and build models. Moreover, if well designed they could also model increasing levels of expectation and the work of forging connections across topics. Since such tasks can take up to several days, they could open up to contain discussions of standards and expectations, including also opportunities to revise in the light of those discussions. In addition, because of certain local options (when, in what order, some details of administration), teachers can become more engaged and more keenly aware of their own role in preparing students in non-routine ways for the assessments. Curriculum-embedded assessments also have the property that can be prepared for in local ways, yet their roughly common format provides teachers across various sites with a common language. Finally, precisely because such tasks would allow us to see what teachers do in leading up to final assessments, curriculum-embedded assessments may provide us with important insights into what contributes to largely shared or widely differential levels of performance among different students.

Curriculum-embedded Assessment in PACE Schools

Since 1990, the Rockefeller Foundation has funded the collaboration of a national network of middle schools to create educational cultures that actually realized access to excellence for all students. That work, known as Project PACE (Performance and Assessment Collaboratives for Education), has had as a part of its agenda, a reconceptualization of assessment as an intrinsic part of instruction Growing out of this work is an extensively different model of student assessment, one that is explicitly designed to insist upon equal access to powerful instruction,

public discussion of excellence (including the possibility of variety), and useful scoring. In PACE schools, we have worked on this issue from two directions: the use of portfolios and the design of curriculum-embedded assessments that can be common-ish across sites. To demonstrate these features, we offer an example from arts and humanities instruction, in which all students were offered access to more than functional literacy. In discussing this example, we want to stress several key characteristics of curriculum-embedded assessments that serve equity concerns. These are: equal access to powerful and sustained instruction, as well as models and public discussion of excellence.

Equal Access to Powerful and Sustained Instruction

Something too easily forgotten in the push to use performance assessments to drive excellence, is that no assessment, no matter how elegant and demanding, can be powerfully equitable, unless all students having access to large ideas and worthwhile strategies. Recognizing this, PACE teachers have deliberately forged assessments to focus teachers' and students' attention not on the familiar and particulate learning objectives, but on key acquisitions during the middle school years.

For instance, PACE teachers have, across diverse classrooms and communities, taken issue with the habitual forms of literacy practiced in many middle school classrooms where serial and atomic reading assignments out of basals, are followed by low-level comprehension questions. Instead, they have decided that middle school literacy must go from emphasizing fluency with familiar materials to "textual power" (Scholes, 1985). This concept has at least four dimensions. The first lies in being able to read a wide variety of texts -- books, song lyrics, visual art, film,

and video -- and integrate information across those diverse sources. The second has to do with being able to read those texts at more than a literal level (e.g., being able to understand imagery, allusion, the stance of an author, the sub-text in what a character has to say). The third aspect is being able to read in the context of a sustained investigation, where the understanding of an issue, a period, or an author builds up, or thickens, over a sustained period of time. The fourth and final feature is that literacy is really about enabling investigation: it is a tool that should make it possible for you to form and follow out a question.

Thus, in one rural classroom, third graders open the year by reading *Wingman*, by Donald Pinkwater, a story of a Chinese-American student making his way in a city and a classroom. Part of what the boy does is to take over the medium of comic illustrations and invent a character, Wingman, who is a miraculous mix of traditional Asian warrior and contemporary American hero. Subsequently, students read extensively in Native American literature, thinking about the manner in which Native peoples developed an imagery reflective of their connection to the natural world: a kinship to animals; an awe of natural forces like changing celestial patterns, storms, and seasons; and an appreciation of the curvalinear patterns of water, wind, and geography. They discuss Native peoples' mythologies as an outgrowth of their relationship to the natural world, using the particular example of the stars, contrasting mythological accounts and contemporary scientific explanations. In this sequence, students also act as authors, observing their own immediate environment, writing poems and chants about the specifics of the natural world they know, and in that way, acquire an understanding of figurative language. In a second, quite different urban classroom, where families have often immigrated and remain highly mobile, teachers have developed quite a different approach, which is appropriate to their students, but no

less demanding. In these classrooms students work on how personal memory, realized in oral and written language, can yield an autobiography and a sense of continuity even across dramatic changes. This is seconded by an additional urban theme: that a person's memories are inevitably colored and shaped by their language and culture. Students and teachers begin by discussing the challenge of making internal experience public via language. Students discuss the various forms that their own memories take: imagery, dreams, family stories, photo albums, even hand-me-down clothing. They narrate some of their own memories, with peers acting first as listeners and then as recorders and/or translators. Students also experiment with writing using "the voice" of some object that has been with their family over years, as a silent to witness to all that has transpired:

Old Kitchen Chair

> They say I am beat up. But I got memory in these old wood bones. I was there, just sitting quietly, when Marta was born, and Martin, and then the little one. I sat through the argument about where to move when the three rooms were to small. I heard Mama and *mi abuela* wanting closer to the family and Tio saying no, closer to the bus line. What they don't know is that I can hear even the little ones whispering, "No, go to the houses near the park." I got as many memories as splinters in these old bones.

In addition, students read widely in other people's accounts, looking at the role of figurative and dramatic language in making the private public. The classroom, in this case, functions much as

a public library, offering children a wide range of choices from illustrated picture books to adult recollections.

Thus, while in no sense following a pre-packaged or identical program, teachers in diverse classrooms have designed curricula, that over a sustained period of time, insist that large numbers of students behave as active readers and writers. In addition, these projects, different as they are, argue that all children should understanding of how many different kinds of texts carry meaning and be able to think about texts as implying history, identity, and values, as well as more familiar and obvious kinds of factual information.

Models and Public Discussions of Excellence

Many assessments simply bid students to show what they know because the test requires it. Such assessments are barren of any models that could either specify or inform performance. In contrast, the assessments, or culminating events, designed by PACE teachers are deliberately designed to illustrate vividly what the qualities of good work are.

At best (which is not always) informal models for good work is everywhere. In fact, the informal discourse that surrounds these culminating events is revealing insofar as it bids students to seek connections and to make use of what they have seen in others' work all along the way. Here, for instance, is a teacher from the rural classroom described above, introducing his students to the actual embedded assessment:

> Do you all remember when we read *Wingman* at the very beginning of the year? Do you remember how Danny used his imagination to keep him company? And

you know how we have been talking about Native Americans' imaginations? Like how they saw stories in the stars? Well, we're about to read about someone else's imagination at work. Someone who lived in the same city as Danny, but who saw it differently....Then later we are going to do something about your own imaginations and you see the world you live in...

Across widely differing preparations (rural sense of place, urban discussions of memory) a culminating event is organized around introducing students to the work of Faith Ringgold, a contemporary African-American artist who creates story quilts where narrative and imagery combine to portray her experience and the history of her community. In this event, students watch a video tape in which Ringgold recounts growing up black, female, poor, and wildly imaginative in New York in the 1930s. She talks boldly and appreciatively about her memories of her mother, a seamstress and fashion designer, and her father, who helped to build the George Washington Bridge. What comes across is Ringgold's zest for her own childhood and the people who taught her "to rise above adversity." Subsequently, children read Ringgold's book *Tar Beach*, in which she talks about the pleasures of going up to the rooftop on summer nights, and her imaginings that she could fly over the bridge, giving it back to her father, who -- though he walked its girders--could not belong to the union, because of his race.

In a second portion of this experience, students have an opportunity to think about what makes *Tar Beach* such a compelling and vivid memory. But they do so in a very active and concentrated way that forces investigation and reflection. They are asked to imagine that they are Faith Ringgold who is revising her book to include more about her mother. They review the

video, taking notes about all that Ringgold says about her mother's influence and help. They choose where to insert their episode but they are asked to do it "seamlessly," attending carefully to making their episode fit into the chronology, the texture, and voice and the imagery of the book. For instance, one student chose a page that shows how Faith (who becomes Cassie Louise Lightfoot in the book) wishes her mother could sleep late, rather than crying while waiting for her father to return empty-handed from job hunting. She wrote:

> Times weren't always fine in our house. When summer went away, taking Tar Beach, we had to live inside. But Mama invented Bed Beach while we waited for Daddy to come home. Bed Beach was a heap of quilts that belonged to Mama--all the ones that came down to her from her mama and her mama's mama. Beebee (Cassie's brother) he like [sic] the Dress Up Quilt, all silky scraps of everybody's party clothes. Me, Cassie Louise Lightfoot, I liked the Mississippi Quilt. It belonged to three families back. To a time when her family came out of slavery. You could lie under it and feel yourself flying through history.

Students read their drafts to one another. They are asked to listen actively, making comments on others' work and keeping their own lists of what they would like to "steal" for their own revisions. Building on their critique and notes, students revise their original first drafts. They openly discuss two questions: "What makes a piece good?" and "What are all the different ways in which pieces can be good?" The first conversation is about standards, the second is about the varieties of excellence (e.g., that goodness can be achieved along any number of routes). What

students uncover often in these discussions are the multiple dimensions of excellence. Strong pieces exhibit sharp language use, imagination, a keen sense for Cassie's spunk, images that come from the urban scene and from African-American culture. But they also realize the multiple ways in which those standards can be realized. Some children write dreams, others create conversations, still others invent new segments of narration. Some are rottenly spelled, but powerful. Others are clear as a bell and flat as pancakes. In the wake of these discussions, students invent informal rubrics, or keys (actually more like brainstormed lists of symptoms to look for). Using these lists, they write critical responses to other students' pieces.

The whole point of this work with Faith Ringgold's *Tar Beach* is to create a clear sense of the resources a writer has to spend and the ways in which a writer can select among those resources to create an effective piece. It is only following these instances of supported performance, and discussions of excellence, that students engage in independent, and unmodelled reading and writing performances in which they make notes, draft, and then revise a piece of their own autobiography. In so doing, they are reminded, over and over again, to raid their fund of resources: what they have read, the energy of Ringgold's talk, and what they could "steal" from what their classmates have invented. (In fact, in one class there is a refrain, borrowed from an art teacher: "The first law of art is if you like it, steal it.") The results are often (though by no means universally) striking. Students do not simply inscribe what they got for their birthday, or the day they lost two teeth at once. They frequently have a keen sense for the resources of language that they can use to drive home the quality of even routine experiences of childhood:

Waiting Up

My mama worked nights. I hated waiting. Some nights, I stood out on the porch in the dark and played a game of listening to footsteps like my ears would break in order to bring her in. It was like my hearing was a big fish hood out to catch her. But once it got dark, early, my granny wasn't having me be out there in the night. So I used to sit inside, leaned up against the window. When a car door slammed, I could feel and hear it rattle. It went though me like an invisible electric wire. "Maybe this time... maybe." I began to listen to all the night sounds- - the horns, and sometimes the sirens, and the lady across the street yelling to her husband not to forget the milk. Then my mama came in the door, and I forgot the old night sounds. It was only her. With the rain rolling quiet off the ends of her boots.

Performance assessments can be helpful in the struggle for equality of educational access insofar as they insist that equal access must extend, not just to a body of inert information, but to the intellectual power to interrogate and use that information. But such modelling is insufficient. As shown here, those assessments must be guided by a clear sense of what the key educational outcomes at given levels of schooling must be. Assessments must grow out of, rather than being independent of, curriculum. That curriculum must approach the teaching of significant outcomes from quite different angles, using local resources, if teachers are actually going to understand, rather than mimic, what it is to provide access. In addition, assessments must function as engaging invitations to join in the pursuit of interesting human challenges pursued by people of all genders, ethnicities and social classes. Even more radically, we believe, based

on our experiences with PACE, that the very body of performance assessments ought to involve students in supported performances and explicit discussions of the dimensions of excellence, *prior to* eliciting independent demonstrations of understanding.

Assessment as an Episode of Learning: The Case of Schools and Systems

But strategies for the equitable design of performance assessments are insufficient. There is a second question: how can we ensure that their *results* will be used to support moves towards greater equity in educational opportunity, as well? The earlier questions focused on avoiding inequitable effects: "How can we avoid making matters worse?" Here we are arguing that we must go one step further, asking "How can we make matters better?"

The great untapped potential of state and district assessment programs is the possibility that they might be used to enable a diagnosis of the educational system and to inform our educational practices. A performance assessment system that gathers data on school resources and school practices as well as student outcomes could suggest links between school inputs, practices, and outcomes which have equity implications, and from which we can derive future policy. In addition, it could encourage reflective discussion among teachers, parents, and others in the community about what constitutes excellence and what curricula and instructional practices are most appropriate and effective in developing students' abilities. In short, an effective performance assessment system may create opportunities for us to learn something about educational practices, and not merely about student outcomes.

This can only be called a hope, or at best, a wager. Given the intellectual and political heritage of assessment policies, it is a hope that can be fulfilled only if such assessments are

designed with this end explicitly and constantly in mind. Even then, the data from performance assessments must be mined to shed light on issues of access to educational opportunity. Such a system requires attention to issues of design, administration, scoring, and reporting. We take these up in this section.

Using Assessments to Understand Inequities

The score gaps between poorly-served and well-served students narrowed during the 1960s and 1970s (Smith & O'Day, forthcoming). If these gains are not real, but due to the fact that we have learned -- unfortunately -- how to simulate equity, at least where those measures are concerned, then one of the benefits of performance assessments may be that they will once again open up to view the uneven playing field of U.S. education. Emerson J. Elliot, U. S. Commissioner of Education Statistics, remarked in a recent talk, for example, that racial differentials are even wider on performance-assessment measures than on traditional standardized tests (Elliot, 1992). Consequently, this is a critical moment to engage in a serious investigation of the nature of educational inequality in the contemporary U.S.

Our response to evidence of wide, and potentially widening, performance gaps hinges on our conclusions about the underlying causes for these differences. If performance assessment grows in popularity, as seems inevitable (for a while at least), and if school performance becomes increasingly important to future earnings (and there is evidence that this has been happening for the last decade (Murnane, Willett, & Levy, in press), then racial, ethnic, and social class differentials on assessment measures will become increasingly detrimental to minority and low-income students' life opportunities. It is imperative, then, that as we develop new assessment

instruments, that we examine carefully the reasons for the performance differentials before using the new instruments in any high-stakes assessment.

There are several flavors of explanations for racial differentials in performance on assessment, each with very different policy implications. If we view a student's performance as a function of three kinds of variables -- those pertaining to the student (biological and, to some extent, cultural factors), those relating to his or her environment (access to educational opportunities, language proficiency, health status, etc.), and those relating to the assessment instrument (design, administration, and scoring) -- we can describe three broad categories of explanation for the observed group performance differences. One or more of these may play a role. Before discussing what is to be done about racial/ethnic/gender/social class performance assessment differentials, we must consider how different explanations will affect the policies and assessment instruments we choose.

The oldest and crudest explanation for group differences in test scores is that performance gaps represent differing levels of innate ability in the domains being assessed. By and large, these kinds of interpretations have been discredited as blatantly racist or sexist (which they generally are).[5] Nonetheless, one often unstated but implicit explanation for continuing racial achievement differences through the 1980s has been the claim that minority and low-income students suffer from some "cultural deficit" -- the lack of motivation, proper upbringing, or caring

[5]They need not be racist or sexist, however. Even if there are heritable genetic differences in "intelligence" or some other characteristic which result in racial test score differentials, however, this does not imply a relationship of inferiority. Judgements about what characteristics are inferior and superior are socially determined, so that test score differentials even between groups that differ genetically may be indicate that tests privilege certain kinds of "intelligence," rather than that certain groups are inferior.

parents. While such explanations differ slightly from those focusing on innate intelligence differences, they are similar in that they blame the performance differentials on some characteristic of the low-achieving group. White/male/middle-class/English-speaking values and characteristics are held as the norm, and minorities, women, and the poor are "deficient" to the extent that they differ.

A second possible explanation for average test score differences between groups is that the differences may reveal the bias of the test toward one set of abilities over another. If, for example, men and women differ innately, on average, in their spatial relations ability, then test score differences between men and women may result from a disproportionate emphasis in the testing instruments on a particular kind of mathematical reasoning (that which predominates in men). Similarly, some assessments may privilege one method of solving problems over others -- if some methods are linked with one gender or one racial/ethnic, cultural, or economic group more than others, this privileging will result in racial score differences. A related form of this bias is cultural bias, where tests or assessments privilege one set of cultural experiences or styles over another.[6]

A different type of test bias occurs when the test is used to make inferences that are less valid for one subgroup of the population than for another. For example, if many immigrant children cannot read English well enough to understand word problems on a math test,

---

[6]See, for example, Sarah Michaels' (1982) study of children's narrative styles. Her study suggests that "children from different cultural backgrounds come to school with different styles and interpretive conventions for using narrative discourse for conversational purposes" (p. 1). In addition, she found that teachers seemed to work better with students whose narrative style matched their own, an effect that "may adversely affect school performance and evaluation" (p. 2) of students from different cultural backgrounds.

interpreting their low average scores as indicating deficiencies in math reasoning skills, or worse, in innate intelligence (as was done with the Army Alpha test), would be erroneous (see Cole & Moss, 1989, for a discussion of this type of bias).

A third potential cause of test score differentials rests in the availability of educational opportunities to learn. Students with less access to quality curricula, instruction, and resources within the school, and to extracurricular educational opportunities outside of school, will naturally perform less well on tests of achievement. In this case, it is not the assessment instrument, but the availability of opportunity to learn which is inequitable; to throw out assessment instruments that reveal evidence of achievement differences may therefore be akin to shooting the messenger who bears unwelcome news.

What we do about performance gaps depends on what we understand about their causes; the remedy must fit the diagnosis. If the performance differences on assessments that Commissioner Elliot refers to are due to test bias, either in the items or the scoring, we must redesign the instruments; if to differences in unequal opportunity, then we must remedy the inequalities that underlie them. The difficulty is that we often have little clear data on to what extent each is an appropriate diagnosis. Most policy debate is framed more by the political realities of the day than by clear evidence of test bias or unequal opportunity. This is not to say, however, that we do not have such data -- in fact, for those who look, there is clear evidence of inequality of educational opportunity and of the importance of opportunity in educational achievement (see, e.g., Epstein & Mac Iver, 1992; Kozol, 1991; Orfield & Reardon, 1992, 1993; Smith & O'Day, forthcoming). The evidence of test bias is less clear, however, because it is harder to prove.

As performance assessment becomes more popular, it will become increasingly necessary to provide clear evidence that the assessments are not culturally biased. Performance assessments may be more open to bias in practice than traditional paper-and-pencil-tests. Or they may be no more biased than multiple-choice tests, but may simply shift the locus of bias from item selection to scoring practices. Performance assessments involve subjective judgements in scoring as well as in design, and this may introduce more opportunities for bias to creep in (of course, the scoring of performance assessments may conversely offer more opportunities to correct for biases in design). The fact is, we know very little about the potential bias lurking in performance assessment scoring systems, nor about how to correct such bias. Either way, performance assessments lack the appearance of objectivity that paper and pencil tests have, and will be therefore all the more open to charges of test bias, regardless of whether they are biased or not.

Consequently, designers of performance assessments and the policy makers implementing them must take particular care to design strategies to distinguish between the effects of test bias and the effects of unequal opportunities. Moreover, they must design assessments in ways that allow them to be used to demonstrate the effects, if any, of unequal opportunity. In other words, assessments can be used as powerful levers for educational equity reform only if they are carefully designed to do so.

One example of a way in which an assessment system could be designed to inform policy and practice comes from Vermont's recent experience with statewide portfolio assessment in fourth- and eighth-grade writing and math. A 1992 RAND report commissioned by the state to evaluate the effectiveness of portfolios as an assessment tool examined the consistency of scores given to the same portfolios by different scorers (Koretz, 1992). The report found a consistent

pattern of low inter-rater reliability, with average reliability coefficients ranging from .33 to .43, but concluded that it could "only hypothesize about the causes of the low reliability" (p. 4). The reason the report could only hypothesize about the causes of the low reliability, however, is not because of some fundamental unknowability of the reasons, but because the program was not designed to inform those implementing it about the reasons for such differences.

The question is particularly important because scoring practice is such an important piece of performance assessment, and because we know so little about it. We know little, for example, about whether scorers of different race, ethnicity, class, or gender differ in the way they score particular portfolios, or portfolios from students of different backgrounds, nor do we know what kind of training is required to reduce such variation. We do not know to what extent teachers from schools with very different populations score differently -- if teachers from high-achieving, privileged schools score differently than those from less advantaged schools, this has important implications for how we design a scoring system. And we do not know whether scores of more proficient students are more or less reliable than those of less proficient students.

For the Vermont program to have been designed to help answer these types of questions, it would have had to have included careful records of who scored which portfolios, detailed background data on the characteristics of the students, teachers/scorers, and schools involved, and data on the instructional and curricular practices of the schools. Such records would inform future training and scoring practices. If, for example, we found that scorers who received no training scored as reliably as those without training, we could conclude that the type of training provided was inadequate. Or if we found that scores given by teachers from advantaged schools were comparable to scores given by teachers from similar schools elsewhere, but lower, for the

same portfolio, than those given by teachers from disadvantaged schools, we could conclude that teachers from different kinds of schools held different definitions of excellence; in such a case, the remedy might include policies of pairing scorers/teachers from advantaged and disadvantaged schools in order to create opportunities for dialogue about standards across districts, something that rarely happens currently.

Assessment as Reflective Practice: The Case of Teachers and Communities

In addition to mining the data on assessments for evidence of unequal access to instruction and curriculum, as well as for evidence of patterns of scoring bias, we want to design ways for assessments to be used to create opportunities for dialogue about educational standards and practices. We want to know how assessment practices can be episodes of learning, not only for the student involved, but for the school and educational system as well. In other words, how do we create structures that encourage teachers, schools, and communities to be reflective about their practices, and to use the results of assessments as opportunities to learn about and improve their practices?

One possible structure is that practiced at Central Park East Secondary School (CPESS) in New York City. There, teachers gather for a thoughtful examination of the portfolio of a student they graduated the previous year as a way of assessing their own practices and the standards to which they are holding their students (McDonald, 1991, 1992). This sort of "self-audit" encourages a reflectiveness within a school community, not only among teachers, but -- if students, parents, and teachers from other schools are included in the audit -- also between the school and its community and among teachers from different schools. One can easily imagine

the conversations: Has this student demonstrated that she has met our educational goals? Would other schools have graduated her as well? If so, what can we learn from the portfolio about our educational practices and our standards of knowledge? And if not, why not? What supports could we have provided to have helped her do better? And why, if she didn't meet the standards, did we graduate her? (cf. Darling-Hammond & Snyder, 1992, pp. 29-34). A somewhat similar practice is developing across PACE sites. There teachers will cross-read the portfolios of other middle school students, in an effort to judge not only whether students meet the standards for strong and capable eighth graders, but whether students have adequate opportunities to learn.

The conversations such practices generate about standards and excellence are potentially very powerful. They allow the school to use assessments as an episode of learning -- as a way of informing future instructional and curricular practices. And they create structures for dialogue within and among schools about the standards and varieties of excellence.

But reflective practices like this depend on several important conditions. They rely, first of all, a society that trusts its teachers, parents, and students to be able to engage in meaningful conversations about educational standards. But we live in a society which is ambivalent at best about its teachers, and less and less willing to give them control over curriculum and instruction, as the recent trends toward centralized student and teacher testing programs and curricular requirements reflect. While much of this mistrust may be unfounded, it is also true that relatively few teachers have the training and experience necessary to engage in the kinds of meaningful dialogues about standards and excellence on which these practices depend. We will need to work hard to develop a cadre of teachers capable of these practices.

Second, reflective practices like those at CPESS imply a system of local control over

standards and assessment, a system that is neither top-down nor bottom-up, but that is characterized primarily by patterns of "horizontal dialogue." If we are serious about holding all students and schools to common, high standards, we need some mechanism for ensuring the consistency of the standards across schools and districts. Minimal competency testing policies, for example, ensure consistency by having some central and objective computer assign scores. But if, on the other hand, we are serious about allowing for varieties of excellence and multiple ways of demonstrating excellence, we must allow for local control. No centralized computer or scoring department will be able to provide the context-rich and instructionally-embedded scoring function that performance assessments require. Thus, we need a system of accountability that is not centralized and bureaucratic, nor one that leaves each school or district to define its own standards of excellence. A system that allows for local autonomy while requiring horizontal dialogue may be the structure most suited to providing this kind of accountability and consistency (for a similar notion of accountability, see Darling-Hammond & Snyder, 1992).

DRAWING THE IMPLICATIONS

In Part I of this paper, we argued that performance assessment practices and policies in the U.S. derive from an historical, philosophical, and political heritage that is, in many ways, antithetical to concerns for equity. They inherit assumptions about intelligence as fixed and about excellence having only a few forms. These are views that gained currency as early as the seventeenth century and they are equally fresh and evident in the concerns of the excellence movement. By virtue of this heritage, performance assessment practices do not come to us in

1993 with any guarantee of their equitable use. If they are to be used as agents of equity, we will have to design them explicitly to do so.

In Part II we discussed a number of promising practices that may encourage the equitable design, administration, and scoring of assessments. Using the particular example of curriculum-embedded assessments, we argued that the key to equity is to treat the assessment practices as opportunities for learning, for schools and teachers as much as for students. By embedding assessment practices in the curriculum, so that assessment becomes an opportunity for students and teachers to discuss standards of excellence (as in the *Plessy* v. *Ferguson* paper example), assessment becomes less a practice of sorting students at the end of the term, and more one of encouraging them to reflect on and improve their work. And by using assessment practices as opportunities to encourage dialogue among parents, teachers, and their communities, as well as among teachers of different schools and communities (as we suggest the CPESS model might do), assessment becomes an opportunity for society to engage in the critical work of rethinking its notions of intelligence, knowledge, and learning.

For us, strong policy implications follow. Some are large and national. Where national assessment is concerned, it would be disastrous if mandated soon or rapidly. The underlying conditions of inequality, and the still open questions about excellence in a highly diverse society, declare against it. In addition, if the results from PACE sites are at all indicative, what makes sense may be a national framework that vigorously requires local realization - possibly at the state or district level. But no district or state should be allowed to enter any such system, without being able to document "equivalent" (albeit not identical) opportunities to learn or without agreeing to interrogate results in the light of a pervasive concern for equity. Finally, if, as

proposed, there is to be a national certifying body that will pass on the integrity of assessments developed by states and projects, that body must actively solicit and judge more than end-of-course assessments. Specifically, it must have the funds to nurture the development of the kinds of curriculum-embedded tasks and sustained projects which we have described here. The hallmark of such assessments is that they scaffold student performances, open up the discussion of the standards for good work, and provide opportunities for revision in the light of new understandings. Such a body should also be empowered to help states and districts to think about tasks that admit of several different, but equally rigorous and useful ways of displaying understanding. Moreover, we must have an alliance with the technical and legal communities to help us think about ways of scoring such complex data that are responsible and fair. (Without their investment, the effort to move away from multiple-choice technologies and the constrained learnings they foster, will be picked apart in technical reviews and court cases.)

But most of this will be written in the language of declaration: "By the year 2000, all states will..." or "By eighth grade, all students should...." But, in truth, students learn in classrooms on hallways in particular schools. It is in face-to-face interaction, and school-level decisions that such declarations will either be realized, trivialized, or ignored. Consequently, for us perhaps the most important (although unfortunately the least glamorous) policy decisions have to do with enabling (at least as much as requiring) schools to act on questions of equity.

Based on our experiences with PACE schools, there are several critical ingredients.

o The re-organization of schools into developmental spans (k-2, 3-5, 6-8, 9-10, 11-graduation). We have substantial research demonstrating that grade-by-grade

retention pre-destines many students to school failure. Thus it is essential to provide periods of developmental time in which teachers, families and students can move towards clearly defined benchmark achievements. School counselling must fall in line: creating sustained conversations with students, and with the adults in their lives about progress toward these benchmarks and the available resources for reaching them.

o The re-organization of the school curricula to concentrate on major capacities (communication and learning, quantitative reasoning, cross-cultural studies) andtechnologies (critical reading, writing, research skills). Schools need to focus their energies and students' attention.

o The re-organization of school time (day, week, and year). We have to move away from forty minute modules to more sustained blocks of time which permit science experiments, play rehearsals, interviews and follow-up discussions. We have to consider alliances with community service, apprenticeships, and cultural organizations so that students have the experiences of applying and transporting their knowledge. There are tough discussions to be had with teachers' unions about work that spans more than 180 days, and resulting contracts that acknowledge that teachers' work involves considerably more than contact hours.

o Substantial support for lateral conversation with other school communities

engaged in realizing standards and the deliberate cross-moderation among schools within a community, and among the schools of different communities. This "horizontal accountability is a necessary addition to the usually recommended top-down (from the state) and bottom-up (individual teachers) approach to setting standards.

o Resources to re-think grouping for learning in schools. We have ample evidence that tracking as we originally invented it has been disastrous. That does not mean that wholesale, all day long heterogeneous grouping is the perfect antidote. We have to think our way towards a solution that would include options for independent work, small group work, tutoring, and whole group instruction.

o Resources for massive professional development. What we are looking at is not a problem of technical re-tooling or picking up the knack of writing performance assessments. We deliberately began this paper by describing the problematic notions of intellect, excellence, and equity that we have inherited. Changing these fundamental beliefs is not a frill for a seminar in values. It represents the most fundamental kind of retraining for many, many current professionals. And this means counsellors, vice-principals and school boards, not "just" teachers.

CONCLUSION

Early in the twentieth century, the Committee of Ten, headed by Harvard President Eliot, intervened in the anarchy of individual college admissions. As early as 1890, Eliot and his colleagues advocated for a single, common set of examinations at a time that admissions had quadrupled and diversified. Originally these were a form of performance tests -- essays written in blue books. However, they were also curriculum-dependent exams that took their content and form almost directly from that of exclusive private schools and the few public exam schools. Wishing to widen the field of candidates dramatically, the Committee argued for curriculum-independent exams that were global investigations into such apparently curriculum-independent fields as vocabulary, analogies, and mathematical problem-solving. When Terman, Yerkes and others developed the Alpha Tests for the US Army during the first World War, a similar multiple choice format was imported to the College Boards, with some advocates even arguing that such a format further democratized entrance by down-playing skills like essay writing (Stewart, in press).

But, any testing technology takes its meaning in context. In the context of radically tracked schools, only some students read enough demanding texts in English to acquire the sampled vocabulary. General math, not algebra and calculus, was the sink-hole for all those thought to be destined (as Terman would have it) for carpentry or offices, rather than laboratories and universities. Consequently, an examination originally entitled the an "aptitude" test could not have been more profoundly a test of what a student was able to achieve -- given his or her

opportunities to learn. Moreover, complementary technologies grew up -- college preparatory classes and college counselling in schools, and cram schools like Princeton Review and Stanley Kaplan appeared at the margins. Their demonstrable ability to raise students' scores by anywhere between one and two hundred points, by teaching "how to work the test," also illustrates how distant an original effort -- despite its intentions -- can come from levelling the playing field. Finally, by modelling multiple choice technologies, an assessment tool, like the College Boards, promulgates a certain deception. Overtly, it suggests that particulate knowledge is what arbitrates entrance, when, in fact, a student's capacity to think (as evidenced in her transcript, outside activities, letters of recommendations and essays) counts at least as much. The effect is to turn public attention (and often instruction) to one end, while leaving untouched those factors (such as access to writing skills) that may have more to do with students' ability to be successful in college.

There are two stiff lessons here: First, any assessment enters a complex, cultural system. Unless we design its uses, we have failed our assignment. The complex social system assessments enter may be indifferent, if not antithetical, to the equitable impulses designed into an assessment system. Thus, we have to be prepared constantly to shift our technologies -- if only to stay one step ahead of forces that would build a stockade around cultural capital.

REFERENCES

Airasian, P. W. (1979). "Educational Measurement and Technological Bases Underlying Minimal Competency Testing." In Peter W. Airasian, George F. Madaus, and Joseph J. Pedulla (Eds.), *Minimal Competency Testing*, 33-47. Englewood Cliffs, NJ: Educational Technology Publications.

Brandt, R. M. (1981). *Public eduction under scrutiny*. Washington: University Press of America.

Britell, J. K. (1980). Competence and excellence: The search for an egalitarian standard, the demand for a universal guarantee. In R. M. Jaeger & C. K. Tittle (Eds.), *Minimum competency achievement testing: Motives, models, measures, and consequences*, pp. 23-39. Berkeley, CA: McCutchan.

Brown v. *Board of Education*. 347 U.S. 483 (1954).

Cannell, J. J. (1987). *Nationally normed elementary achievement testing in America's public schools*. Daniels, WV: Friends for Education.

Carstens, L. (no date). Social and political influences on the American testing system: a study of equity issues. Unpublished manuscript.

Cheney, L. (1991). *National tests: What other countries expect their students to know*. Washington, D.C.: National Endowment for the Humanities.

Cohen, D. K., & Haney, W. (1980). Minimums, competency testing, and social policy. In R. M. Jaeger & C. K. Tittle (Eds.), *Minimum competency achievement testing: Motives, models, measures, and consequences*, pp. 5-22. Berkeley, CA: McCutchan.

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.) (pp. 201-219). New York: American Council on Education and Macmillan Publishing Company.

Crane, J. (1991). Effects of neighborhoods on dropping out and teenage childbearing. In C. Jencks and P. Peterson (Eds.), *The urban underclass*, 299-320. Washington, DC: The Brookings Institute.

Cuban, L. (1984). *How teachers taught*. New York: Longman.

Darling-Hammond, L., & Snyder, J. (1992). Reforming accountability: Creating learning-centered schools. In Ann Lieberman (Ed.), *The changing context of teaching*, pp. 11-36.

Chicago: University of Chicago Press.

Digest of Education Statistics: 1985-86. (1986). W. V. Grant & T. D. Snyder, (Eds.), Office of Educational Research and Improvement, U.S. Department of Education Center For Statistics. Washington, DC: U.S. Government Printing Office.

Epstein, J. L., & Mac Iver, D. J. (1992) *Opportunities to learn: Effects on eighth graders of curriculum offerings and instructional approaches.* Report No. 34, Center for Research on Effective Schooling for Disadvantaged Students, Johns Hopkins University.

Farrar, E. (1990). "Reflections on the first wave of reform: reordering America's educational priorities." In Stephen L. Jacobson & James A. Conway (Eds.), *Educational Leadership in an Age of Reform*, pp. 3-13. New York: Longman.

Fass, P. (1989). (check title and publisher).

Goodlad, J. & Keating, P. (1990). *Access to knowledge.* New York: The College Entrance Examination Board.

Gould, S. J. (1981). *The mismeasure of man.* New York: W. W. Norton & Co.

Graham, P. A. (1993). What America has expected of its schools over the past century. *American journal of education, 101*(2), p. 83-98.

Greene, M. (1980). Response to "Competence and excellence: The search for an egalitarian standard, the demand for a universal guarantee," by Jenne K. Britell. In R. M. Jaeger & C. K. Tittle (Eds.), *Minimum competency achievement testing: Motives, models, measures, and consequences*, pp. 40-48. Berkeley, CA: McCutchan.

Haney, W., & Madaus, G. F. (1979). Making Sense of the Competency Testing Movement." In Peter W. Airasian, George F. Madaus, and Joseph J. Pedulla (Eds.), *Minimal Competency Testing*, 49-72. Englewood Cliffs, NJ: Educational Technology Publications.

Jensen, A. R. (1969). How much can we boost IQ and scholastic achievement? *Harvard Educational Review* 39/1 (Winter): 1-123.

Kamin, L. J. (1974). *The science and politics of I.Q.* New York: John Wiley & Sons.

Kluger, R. (1975). *Simple justice: The history of* Brown v. Board of Education *and black America's struggle for equality.* New York: Vintage.

Koretz, D. (1992). The reliability of scores from the 1992 Vermont portfolio assessment program, interim report, December 4, 1992. Washington, DC: RAND Institute on

Education and Training.

Kozol, J. (1991). *Savage inequalities: Children in America's Schools.* New York: Crown.

Liebman, J. S. (1990). Implementing *Brown* in the nineties: Political reconstruction, liberal recollection, and litigatively enforced legislative reform. *Virginia Law Review, 76,* pp. 349-435.

Madaus, G., & Kellaghan, T. (1992). British experience with "authentic" testing. Unpublished manuscript.

Mayer, S. E. (1991). How much does a high school's racial and socioeconomic mix affect graduation and teenage fertility rates? In C. Jencks and P. Peterson (Eds.), *The urban underclass*, 321-341. Washington, DC: The Brookings Institute.

McDonald, J. P. (1991). Three pictures of an exhibition: Warm, cool, and hard. Studies on Exhibitions (No. 1). Providence, RI: The Coalition of Essential Schools.

McDonald, J. P. (1992, December). Talk at Harvard Graduate School of Education.

Michaels, S. (1982). *"Sharing time": Children's narrative styles and differential access to literacy.* Dissertation. Berkeley, CA: University of California, Berkeley.

Murnane, R. J., Willett, J. B., & Levy, F. (in press). The growing importance of cognitive skills in wage determination.

Orfield, G., & Reardon, S. F. (1992, September) Separate and unequal schools: Political change and the shrinking agenda of urban school reform. Paper prepared for the American Political Science Association Annual Meeting, Chicago, IL.

Orfield, G., & Reardon, S. F. (1993, in press) Race, poverty, and inequality. In S. M. Liss & W. L. Taylor (Eds.), *New opportunities: Civil rights at a crossroads.* Washington, DC: Citizens' Commission on Civil Rights.

National Educational Goals Panel. (1991). ...

National Council on Education Standards and Testing. (1992). *Raising standards for American Education.* Washington, D.C.: U.S. Government Printing Office.

New Standards Project. (1991). *A Summary of the New Standards Project.* Unpublished manuscript.

Oakes, J. (1985). *Keeping track: How schools structure inequality.* New Haven: Yale

University Press.

Pedulla, J. J., & Reidy, E. F. Jr. (1979). "The Rise of the Minimal Competency Testing Movement." In Peter W. Airasian, George F. Madaus, and Joseph J. Pedulla (Eds.), *Minimal Competency Testing*, 23-32. Englewood Cliffs, NJ: Educational Technology Publications.

Plessy v. *Ferguson*. 163 U.S. 537 (1896).

Reed v. *Rhodes*. 455 F. Supp. 546 (1978).

Resnick, L. B. (1987). Learning in school and out. *Educational Researcher*, (December), pp. 13-20.

Scholes, R. (1985). *Textual power*. New Haven, Ct.: Yale University Press.

Scribner, S. (1984). Studying working intelligence. In B. Rogoff and J. Lave (Eds.), *Everyday cognition*. Cambridge, MA.: Harvard University Press.

Sizer, T. (1984) *Horace's compromise: The dilemma of the American high school*. Boston, MA: Houghton Mifflin.

Smith, J. A., & O'Day, M. S. (forthcoming). Systemic school reform and educational opportunity. In S. Fuhrman (Ed.), *Designing coherent education policy: Improving the system*. San Francisco: Jossey-Bass.

Stell v. *Savannah-Chatham Board of Education*. 220 F. Supp. 667 (1963).

Stewart, D. M. (in press). The evolution of College Entrance Examinations. In J. Baron and D. Palmer Wolf (Eds.) The National Society for the Study of Education Yearbook (1994).

Terman, L. M. (1922). The great conspiracy. *New Republic*, pp. 116-120.

Toulmin, S. (1987). *Cosmopolis*. Chicago, Il.: University of Chicago Press.

Wiggins, G. (1989a). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan*, 70, pp. 703-713.

Wiggins, G. (1989b). Questions and answers on authentic assessment. Paper presented at Beyond the Bubble: Curriculum Assessment/Alignment Conference, Sacramento, CA.

Wolf, D. (1989). Portfolio assessment: Sampling student work. *Educational Leadership*, *46* (7), 35-39.

Wolf, D. (in press). *Presence of mind, performance of thought.* New York: College Entrance Examination Board.

Wolf, D. (1993). A framework for English 12 in Pacesetter. New York: The College Entrance Examination Board. Unpublished manuscript.

Wolf, D. & Baron, J. B. (in press). Options for national assessment. In D. Stevenson & E. Baker (Eds.) Performance Assessment (tentative title). Hillsdale, N. J.: Lawrence Erlbaum Associates.

Wolf, D., Bixby, J., Glenn J. III., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. In Gerald Grant (Ed.), *Review of research in education* 17, pp. 31-74.

FIGURE 1: TEACHER'S INITIAL COMMENTS TO THE STUDENT (INCLUDING STUDENT UNDERLINING AND MARKING, KEYED TO HIS RESPONSE

Plessy v Ferguson

You have done an outstanding job of examining the arguments presented in the Plessy v Ferguson Supreme Court ruling, especially those put forth by the plaintiff. In the process, you have clearly learned a great deal about the law, both constitutional and state, and about societal conditions which can either reinforce or weaken the law itself.

You met all the deadlines required in the assignment. The outline clearly established the focus of your paper. Notecards reveal extensive research. Form is nearly flawless — title page, bibliography and footnoting. The title itself is rather awkward. You make excellent use of source available, particularly in regard to the arguments put forth by the plaintiffs and, to some degree, the conditions in the South prior to the Plessy ruling.

Use of quotations adds considerably to the quality of your paper, especially the arguments our forth by the law, representing Plessy. The quotations are always effectively presented.

You have worked hard to establish your thesis, examining the foundations existing in the South for the ultimate weakening if not ignoring the 14th Amendment. You present the briefs in the Supreme Court case with clarity and understanding. I wasn't sure why you labeled the plaintiff's arguments as vague and lacking clarity. It seemed to me that the points were clear, but as you said, in your conclusion, they in themselves, could not persuade the court that the Jim Crow law violated the 14th amendment.

2.

I found the introduction somewhat confusing. Although it clearly show that the interpretation of the 14th amendment was uncertain, it does not give much of an indication of what the body of your paper is all about.

The details about the status of the law and of the social practices in the South showed an impressive understanding on your part of the vast difference between law in the South and equal rights. It was also interesting to know that cities such as Columbia were not nearly as rigid as others.

3. Your paper focuses to too great a degree in the plaintiff's arguments in Plessy. Though they have bearing on the rest of your paper, the arguments take on a place in your writing that seems inordinately large. The arguments of the defense are dealt with in a page, and the actual ruling is described in a rather cursory way. The dissenting opinion merits only a sentence, despite its sensitivity to the arguments put forth at great length in your paper by the plaintiffs.

4.

Thus, it seems that as scholarly as your approach is in this research paper, there was not enough attention on your part to creating a balanced examination which would give your readers more of an idea of the dimensions of the controversy.

5

Your paper is well written. You write clearly, presenting your ideas in an organized manner. The documentation of your research is always in evidence.

GRADE: B+

FIGURE 2: STUDENT'S RESPONSE TO TEACHER'S COMMENTS (NUMBERING REFERS TO NUMBERED AND UNDERLINED SECTIONS IN THE TEACHER'S REMARKS)

3/30/90

Mr.

I appreciate your concern, and willingness to give a second look. I'll try to address points in the order that you wrote them in your comment. I suggest you read my underlining, then read my reply. → on your comments

1. I believe the title is accurate, and helps to tell the reader that this paper will focus on the origins and arguments of the Plessy Vs. Ferguson case. I believe that I focused quite clearly on both, explaining the origins with 6 pages committed to the social and legal history of the era. I documented both positive and negative conditions in the south, from a number of different sources. The arguments are not presented in their entirety, and they are taken from one source, yet I believe I fully explained the significance of these arguments (not found in the sources.)

2. I labeled the Plaintiff's arguments as vague and lacking in clarity. Perhaps it would have been more clear if I had said that the plaintiff's arguments were vague and lacking in clarity before the law. Many of the arguments seemed to push legal thought, and at times the arguments against the law seemed like tangents, failing to mesh into a main idea. The word "vague" may not have been the best choice, but Plessy's arguments were not conventional, an important point to be noted.

3. To have given equal weight to both Plessy's and the states' arguments, in my mind, would have been ridiculous. The idea that a black man was to challenge a white man's law in the supreme court in the late 19th century, that in itself should be a clue that the black man's arguments must be stupendous and overwhelming, if he seeks to win. Furthermore, the burden of proof was on Plessy, not the state. The state was content to stick by its law, whereas Plessy had to produce the proof that this law was unconstitutional. Therefore, Tourgee and his associates used virtually everything they could, the 14th amendment argument, the damaged property argument, and the precedent Brown case, to name a few. The bottom line is that this case, in my opinion, simply cannot be presented as an equally argued affair. It wasn't, and that's what makes it Plessy Vs. Ferguson.

You write, "the (plaintiff's) arguments take on a place in your writing that seems inordinately large." I believe the plaintiff's arguments were inordinately large. Tourgee and his associates present these sweeping, grand, unconventional arguments. Meanwhile the defense simply defends the law as a social necessity that is constitutional. Tourgee's digging into every pocket, creating inordinately large arguments in hopes of swaying the court. He asks the court to interpret the 14th amendment through the eyes of its drafters. Tourgee has no proof, yet he must provide overwhelming proof to sway the court.

4. The reason the ruling is dealt with so swiftly is twofold.

1. I simply wasn't focusing on the ruling. I only hoped to provide a quick overview to tell the reader how it turned out.

FIGURE 2: CONTINUED

2. By the time I reached the ruling, I was on page 16. Knowing that I shouldn't ramble on, I mentioned the decision, and a brief opinion on why the court ruled against Plessy.

5. I'm not sure what a balanced examination of Plessy would have yielded, and I know I did not examine the trial from start to finish. I think it is important for you to know that the book I had that spanned the entire case, was over 200 pages. I focused on what I wanted to focus on. I picked out of the trial what I thought was really interesting, and I think I went on to describe the origins and arguments of Plessy in an in depth, helpful, and interesting way.

Now here's what I feel. I spent a good deal of time thinking out the thesis of this paper. I developed two main ideas;

1. The importance of showing the reader the checker board of Jim Crow laws in the south. I wanted the reader to see that the south was not uniform in its approach, and that the Plessy case gave the south its legal doctrine for segregation.

2. The reader deserved to see the amazing work of Tourgee and the other lawyers. I wanted to show the readers Tourgee's ideas, and to explain why Tourgee lost.

In my opinion these are interesting ideas, and I think I explained them well, and supported my thesis. I guess my concern is that my hard history work suffered because it was not a complete work.

Thanks,

FIGURE 3: TEACHER'S REPLY TO STUDENT'S RESPONSE

Dear _____

You have done a superb job of defending your position. Your argument's are generally persuasive and always articulate and forthright. I feel that you have given me a much clearer idea of what the objectives of your paper were. I wish those objectives had been more clearly stated in your introduction. In any case, I accept your arguments for placing emphasis on Plessy's suit. I also find your concerns compelling regarding your efforts. I don't think I gave adequate consideration to the scholarship involved in your research and writing.

I understand the problem of length, but it does seem essential that you give some consideration to the opposition's dissent — its connection to Tourgee's arguments. You would have had to cut out some of your statements to make the paper meet out length requirements, but it would have been worth it.

A—