Education Policy Analysis Archives

Volume 6 Number 13 **July 14, 1998** **ISSN 1068-2341**

A peer-reviewed scholarly electronic journal.
Editor: Gene V Glass Glass@ASU.EDU.
College of Education
Arizona State University,Tempe AZ 85287-2411

Consequences of Assessment: What is the Evidence?

William A. Mehrens
Michigan State University

Abstract Attention is here directed toward the prevalence of large scale assessments (focusing primarily on state assessments). I examine the purposes of these assessment programs; enumerate both potential dangers and benefits of such assessments; investigate what the research evidence says about assessment consequences (including a discussion of the quality of the evidence); discuss how to evaluate whether the consequences are good or bad; present some ideas about what variables may influence the probabilities for good or bad consequences; and present some tentative conclusions about the whole issue of the consequences of assessment and the amount of evidence available and needed.

I. INTRODUCTION

It is a pleasure to address friends, colleagues, and associates on what I believe to be an important topic -- what evidence do we have regarding the consequences of assessment. I actually chose this topic at last year's (1997) convention when I attended a symposium on consequential validity. As most of you probably know, I am not a fan of the *term* "consequential validity." However, I am interested in the consequences of assessment, and I hope all of you are also. Last year's symposium had such illustrious speakers as Ross Green, Suzanne Lane, Bob Linn, Pam Moss, Mark Reckase, and Elizabeth Taleporos. It was a great session. While they agreed on many things, I perceived some differences in opinion about the amount, quality and interpretation of the evidence regarding the consequences of assessment. I left that session believing that not enough evidence was available but that it would be worthwhile to review the evidence more thoroughly. Then, last summer (1997), at the Council of Chief State School Officers Large Scale Assessment Conference, Peter Behuniak, Bob Linn, David Miller,

and Gloria Turner presented evidence they had regarding the consequences of assessment. While I was very impressed with their scholarship, I again was left feeling it would be worthwhile to investigate the topic further. In addition to the fact that the scholarly presentations mentioned above left me unsatisfied with respect to the evidence on consequences, there are additional rationales for choosing this topic.

Many, but certainly not all, political leaders at the national, state and local levels have been touting the value of large scale assessment. For example, President Clinton and Secretary of Education Riley have argued that voluntary national tests of reading at grade four and mathematics at grade eight would have positive consequences for education. Secretary Riley has said, "I believe these tests are absolutely essential for the future of American education" (Riley, 1997a, quoted from Jones, 1997, p.3) (Note 2). President Clinton has also asked each state to adopt tough standards for achievement. The argument seems to be that if tough standards are adopted, achievement will rise.

Educational reformers suggest that:

> Assessments play a pivotal role in standards-led reform, by: communicating the goals. ... providing targets..., and shaping the performance of educators and students. Coupled with appropriate incentives and/or sanctions--external or self-directed--assessments can motivate students to learn better, teachers to teach better, and schools to be more educationally effective. (Linn and Herman, 1997, iii).

Note the word "can" in the above quotation. The question is, do they? Linn, Baker and Dunbar (1991) pointed out that it cannot just be assumed that a more "authentic" assessment will result in better classroom activities. Linn also correctly suggested that:

> Evidence is also needed that the uses and interpretations are contributing to enhanced student achievement and at the same time, not producing unintended negative outcomes (1994, p. 8).

There is no question that assessment is perceived by many as having a potential for good in both the evaluation of the schools and, if believed necessary, the reforming of them. But there are reasons to question whether that potential will be realized. As Goodling has stated,

> If testing is the answer to our educational problems, it would have solved them a long time ago. American students are tested, tested, tested, and the Clinton administration is proposing to test our children again (August 13, 1997).

Goodling suggested that thinking that new tests will lead to better students is "akin to claiming that better speedometers make for faster cars." (quoted from Froomkin, 1997).

There are both potential values and potential dangers in large scale testing. Is the potential value of assessment a vision or an illusion? Are the potential dangers likely to be realized? What is the evidence?

For testing to be a good thing, the positive consequences must outweigh the negative consequences -- by some factor greater than the costs. The costs of large scale assessments are particularly high for alternative forms of assessment such as have been used in Kentucky. Are the consequences of assessment worth the cost? Are the consequences of alternative assessments worth the much greater cost? What is the evidence?

General Overview

In this presentation I wish to spend a brief amount of time on the prevalence of large scale assessments (focusing primarily on state assessments); discuss the purported purposes of these assessment programs; enumerate some potential dangers and potential benefits of such assessments; investigate what the research evidence says (and does not say) about assessment consequences (including a discussion of the quality of the evidence); discuss how to evaluate whether the consequences are good or bad; present some ideas about what variables may influence the probabilities for good or bad consequences; and present some tentative conclusions about the whole issue of the consequences of assessment and the amount of evidence available and needed.

Because the evidence is insufficient, my tentative conclusions about the consequences of assessment will, at times, obviously and necessarily be based on less than adequate evidence. It may seem drawing such conclusions runs counter to a general value of educational researchers -- that inferences should be based on evidence. I am firmly on the side that more evidence is needed and that inferences should be drawn from such evidence. I am firmly against passing pure proselytizing off as if it is research based. I am not opposed to drawing tentative conclusions from less than perfect data. But crossing the line from basing inferences on evidence to basing inferences on a will to believe should not be done surreptitiously, and I try hard to avoid that in this paper.

For those of you who do not make it to the end, I will tell you now that the conclusion will be that the evidence is reasonably scarce (at least insufficient), and equivocal.

II. POPULARITY, PREVALENCE, PURPOSES AND FORMAT OF LARGE SCALE ASSESSMENT PROGRAMS

A. Popularity

Large scale assessment programs are, at the abstract level, popular with politicians and the public. This is true for both proposed and actual state level assessments and the proposed national level "voluntary" assessments. One example of the popularity is obtained from the 29th Gallup Poll. That poll showed that 57% of the public favor President Clinton's proposed voluntary national test (Rose, et al., 1997). However, when proposals get more specific, there can be opposition -- as witnessed by the opposition of many groups after the proposed testing plan got more specific (e.g., with respect to what languages the test would be administered in). It should be noted that front line educators -- those that might be most informed about the potential value of the proposed voluntary national test are far less favorable than the general public. Langdon (1997), reporting on the PDK poll of teachers found that 69% are opposed to Clinton's proposal. Measurement specialists might also have a reasonable claim to being more informed than the politicians or the public. The comments that appeared on the Division D listserve (http://aera.net/resource/) suggest that there are far more negative views among listserve authors about the value of such tests than there are positive views.

B. Prevalence

Regarding prevalence, state programs have been prevalent for at least fourteen years. As early as 1984, Frank Womer stated that "clearly the action in testing and assessment is in state departments of education" (Womer, 1984, p. 3). In identifying reasons for this, Womer stated that:

> Lay persons and legislators who control education see testing- assessment as a panacea for solving our concerns about excellence in education (Womer, 1984, p.3).

Anderson and Pipho reported that in 1984, 40 states were actively pursuing some form of minimum competence testing (1984).

Of course, it turned out that these minimum competency tests were not a "panacea" for concerns about educational excellence, although there exists some debate about whether they were, in general, a positive or negative force in education. In general, there has been a change from testing what were called minimum competencies to testing what we might now call world class standards. And, there has been a bit of a change in how we assess -- with a movement away from sole reliance on multiple- choice tests to the use of alternative forms of assessment. The most recent survey of trends in statewide student assessment programs (Roeber, Bond, and Braskamp, 1997) reveals that 46 of the 50 states have some type of statewide assessment.

C. Purposes and stakes of state assessment programs

The two most popular purposes for assessment according to respondents to a survey of state assessment practices were the "improvement of instruction" -- which was mentioned by 43 states, and "program evaluation" -- mentioned by 38 states. Some other reported purposes (stated in order of number of states mentioning the purpose) include school performance reporting (33), student diagnosis (27), high school exit requirement (17), and school accreditation (11). (Roeber et al., 1997).

However, measurement experts have suggested for some time that "tests used primarily for curriculum advancement will look very different from those used for accountability" (Anderson, 1985, p. 24) and they will have different intended and actual impacts. Likewise, tests used for high stakes decisions (e.g. high school graduation and merit pay) are likely to have different impacts than those used for low stakes decisions (e.g. planning specific classroom interventions for individual students).

It will not always be possible to keep purpose and stakes issues separated when discussing consequences. However, when evidence (or conjecture) about consequences applies to only a specific purpose or level of stakes I will try to make that clear.

D. Format questions

A fairly hot issue in recent years is whether the format of the assessment should vary depending on purposes and whether assessments using different formats have different consequences. In the Roeber et al., (1997) survey of states, multiple choice items were used by 41 states, extended response item types were used by 36, short written response by 24, examples of student work by 10, and what was labeled as performance assessments were used by 9 states. Four states used what were termed projects.

Performance assessment advocates have claimed that the format is important and that positive consequences come from such a format and negative consequences come from multiple-choice formats. Others, like me, are less sure of either of these positions. As with purpose and stakes issues, it will not be possible to always keep these format issues separated in this presentation, but attempts will be made.

Because in recent years more positive claims have been made for "performance assessments" many of the recent attempts to gather consequential evidence have been

based on assessments that have used performance assessments. However, I tend to agree with Haney and Madaus when they suggested "...what technology of assessment is used probably makes far less difference than how it is used" (1989, p. 687).

III. EVIDENCE ON THE CONSEQUENCES OF ASSESSMENT PROGRAMS

Lane (1997) developed a comprehensive framework for evaluating the consequences of assessment programs -- concentrating primarily on performance-based assessments. She suggested that both the negative and positive consequences need to be addressed and that one needs to consider both intended and plausible unintended consequences.

For purposes of this presentation, I will discuss some major potential benefits and dangers as follows:

- A. Curricular and instructional reform: Good, bad, or nonexistent?
- B. Motivation/morale/stress/ethical behavior of teachers: Increase of decrease?
- C. Motivation and self-concepts of students: Up or down?
- D. True improvement in student learning, or just higher test scores?
- E. Restore public confidence or provide data for critics?

Evidence on consequences is somewhat sketchy, but the "Lansing State Journal" did report one result in big headlines: "Test Results make School Chief Smile" (Mayes, 1997, p. 1) Many of you have seen such headlines in your own states. When scores go up, the administrators are happy and act as if that means achievement has gone up. It may have, but note well that the consequence I am reporting here (tongue in cheek) is that the superintendent smiled -- not that achievement had improved!

In general there is much more rhetoric than evidence about the consequences of assessment and "too often policy debates emphasize only one side or the other of the testing effects coin" (Madaus, 1991, p. 228). Baker et al., in an article on policy and validity for performance based assessment reported that "less than 5% of the literature cited empirical data." (1993, p. 1213). As they pointed out,

> Most of the arguments in favor of performance-based assessment ... are based on single instances, essentially hand-crafted exercises whose virtues are assumed because they have been developed by teachers or because they are thought to model good instructional practice. (Baker et al., 1993, p. 1211).

I would conclude, as Baker et al. did, that "a better research base is needed to evaluate the degree to which newly developed assessments fulfill expectations" (1993, p. 1216). Koretz suggested that:

> Despite the long history of assessment-based accountability, hard evidence about its effects is surprising sparse, and the little evidence that is available is not encouraging. ...The large positive effects assumed by advocates...are often not substantiated by hard evidence.... (Koretz, 1996, p. 172.).

Reckase (1997) pointed out one of the logical problems in obtaining evidence on the consequences of assessments. The definition of a consequence implies a cause and effect relationship, but most of the evidence has not been gathered in a manner that permits a scholar (or anyone else with common sense) to draw a causative inference.

Green (1997) mentioned many problems in doing research on the consequences of assessment. Among them are that few school systems will welcome reports of unanticipated negative consequences, so cooperation may be hard to obtain; there will be disagreements about the appropriate criterion measures of the consequences; cause-effect conclusions will be disputed; and much of the research undertaken is likely to be undertaken by those trying to prove that what exists is inferior to their new and better idea.

Much of the research has been based on survey information from teachers and principals and, as many authors have pointed out, classroom observations might be more compelling information (see, for example, Linn, 1993 and Pomplun, 1997). Research by McDonnell and Choisser (1997) employed three data sources: face to face interviews, telephone interviews, and assignments collected from the teachers along with a one-page log for each day in a two week period. As the authors pointed out, instructional artifacts is a relatively new strategy --not likely as good as classroom observations, but not as expensive either.

Although evidence is sketchy, there is some! I will discuss such evidence under the headings given earlier regarding the possible dangers and potential benefits of assessment programs.

A. Curricular and instructional reform: Good, bad or nonexistent?

Curricular and instructional reform typically means changing the content of the curriculum or the process of instruction. Not quite fitting either of those categories is changing the length of the school day or the school year. Pipho (1997) has reported that one change between the first and second year of state assessments in Minnesota was the addition of summer school offerings, Saturday classes, and after- school remedial programs. That kind of "reform" is mentioned at other places in the literature, and is, at least arguably, a valuable consequence.

With respect to the more traditional meanings of curricular and instructional reform, it has been commonly assumed that assessments (at least high stakes assessments) will influence curriculum and instruction. One often hears the mantra that "WHAT YOU TEST IS WHAT YOU GET." Taleporos stated flatly at last year's AERA session that "we all know that how you test is how it gets taught." (Taleporos, 1997, p. 1). Actually, the evidence for a test's influence on either curricular content or instructional process is not totally clear. And it will vary by the stakes. Porter, Floden, Freeman, Schmidt, and Schwille reported more than ten years ago that:

> Another myth exposed as being only a half truth is that teachers teach topics that are tested. Little evidence exists to support the supposition that national norm-referenced, standardized tests administered once a year have any important influence on teachers' content decisions. (Porter, et al., 1986, p. 11).

But the "myth" persists. Is it a half truth, a full truth, or just wrong? Logic suggests it may depend on stakes, rigor of the standards, and just what the content is (Airasian, 1988). Some anecdotal evidence also supports the importance of stakes. For example, Floden (personal communication, 1998) states that while in the Content Determinants work (the Porter et al. study just cited) few teachers paid attention to the tests, he is now working in districts where losing accreditation is a real threat, so teachers are busy aligning curriculum to the Michigan Educational Assessment Program (MEAP) and setting aside time before the test for MEAP-specific work.

1. Impact of multiple-choice minimum competency tests on curriculum and instruction

Minimum competency tests were not primarily designed to "reform" the curriculum. Rather, they supposedly measured what schools were already teaching. The tests were intended to find out whether students had learned that material and, if not, to serve as motivators for both the students and the educators. The intended curricular/instructional effect was to concentrate more on the instruction of what was considered to be very important educational goals.

Some earlier writings on the impact of multiple-choice tests suggested that the tests resulted in teachers narrowing the curriculum and corrupting teaching because teachers turned to simply passing out multiple-choice question work sheets. The critics argued that education was harmed due to the narrowing of the curriculum and the teaching and testing for only low level facts.

Aside from the confusion of test format with test content (true measurement experts realize that multiple choice tests are not limited to testing facts), there is insufficient evidence to allow any firm conclusion that such tests have had harmful effects on curriculum and instruction. In fact, there is some evidence to the contrary. Kuhs, Porter, Floden, Freeman, Schmidt, and Schwille (1985) reported that:

> the teachers' topic selection did not seem to be much influenced by the state minimum competencies test or the district-used standardized tests (Kuhs, et al., 1985, p. 151).

There is no evidence of which I am aware showing that fewer high level math courses are taught (or taught to fewer students) in states where students must pass a low level math test in order to receive a high school diploma.

There are a few studies (quoted over and over again) which presumably show that elementary teachers align instruction with the content of basic skills tests (e.g. Madaus, West, Harmon, Lomax and Viator, 1992; Shepard, 1991). And I believe those studies have some validity. It is hard to believe that tests with some stakes connected to them will not have some influence on curriculum and instruction. For example, Smith and Rottenberg (1991) reported on "an extensive research study." "The consequences of external testing were inferred from an analysis of the meanings held by participants and direct observation of testing activities..." (p. 7). They concluded, among other things that (1) external testing reduces the time available for ordinary instruction, (2) testing affects what elementary schools teach -- in high stakes environments, schools neglect material that external tests exclude, (3) external testing encourages use of instructional methods that resemble tests, and (4) "as teachers take more time for test preparation and align instruction more closely with content and format, they diminish the range of instructional goals and activities" (1991, p. 11).

Thus, there are studies suggesting that multiple-choice tests result in a

narrowing of the curriculum and more drill work in teaching. But, in fact, the studies are few in number and critics of traditional basic skills testing accept the studies somewhat uncritically. In my opinion, the evidence is not as strong as the rhetoric of those reporting the research would suggest and there is some research evidence that teachers do not choose topics based on the test content (Kuhs et al., 1985).

Green (1997) discussed the evidence and questioned the conclusion that multiple choice tests are harmful stating that "I believe that the data just cited opens to question the assertions about the evils of multiple-choice tests." (Green, 1997, p. 4).

2. Impact of performance assessments on curriculum and instruction

Much of the recent research and rhetoric has been concerned with the effects of performance assessments. Performance assessments are popular in part just because of their supposed positive influence on curricular and instructional reform. Advocates of performance assessment treat as an established fact the position that teaching to traditional standardized tests has "resulted in a distortion of the curriculum for many students, narrowing it to basic, low-level skills" (Herman, Klein, Heath, and Wakai, 1994, p. 1). Further, professional educators have been pushing for curricular reform, suggesting that previous curricula were inadequate and, generally, focused too much on the basics. The new assessments should be more rigorous and schools should be held responsible for these more rigorous standards. As a SouthEastern Regional Vision for Education (SERVE) document entitled "A new framework for state accountability systems" (September 8, 1994) pointed out, some legislative initiatives

> ignored a basic reality: Those schools that had failed to meet older, less rigorous standards were no more able to meet higher standards when the accountability bar was raised. As a result, state after state is confronted with previously failing schools failing the new systems (SERVE, 1994, p. 2).

What does the research tell us about the curricular and instructional effects of performance assessments? Khattri, Kane, and Reeve visited sixteen schools across the United States that were developing and implementing performance assessments. They interviewed school personnel, students, parents and school board members; collected student work; and conducted observations. They concluded that:

> In general, our findings show that the effect of assessments on the *curriculum* teachers use in their classrooms has been marginal, although the impact on *instruction* and on *teacher roles* in some cases has been substantial (Khattri, Kane, and Reeve, 1995, p. 80).

Chudowsky and Behuniak (1997) used teacher focus groups from seven schools representing a cross section of schools in Connecticut. These focus groups discussed their perceptions of the impact of the Connecticut Academic Performance Test -- an assessment that uses multiple-choice, grid-in, short answer and extended response items. Teachers in all but one

of the schools reported that preparing students and aligning their instruction to the test "resulted in a narrowing of the curriculum" (Chudowsky and Behuniak, 1997, p. 8). Regarding instructional changes, "teachers most frequently reported having students 'practice' for the test on CAPT sample items" (p. 6). However the schools also reported using strategies "to move beyond direct test preparation into instructional approaches" (p. 6). Teachers also

> consistently reported that the most negative impact of the test is that it detracts significantly from instructional time. Teachers at all of the schools complained vehemently about the amount of instructional time lost to administer the test (p. 7).

Koretz, Mitchell, Barron, and Keith (1996) surveyed teachers and principals in two of the three grades in which the Maryland School Performance Assessment Program (MSPAP) is administered. As they reported, the MSPAP program "is designed to induce fundamental changes in instruction" (p. vii). While about three-fourths of the principals and half of the teachers expressed general support for MSPAP, fifteen percent of the principals and 35% of the teachers expressed opposition. One interesting finding was that about 40% of fifth-grade teachers "strongly agreed that MSPAP includes developmentally inappropriate tasks" (p. viii). One of the summary statements made by Koretz Mitchell, Barron, and Keith (1996) is as follows:

> The results reported here suggest that the program has met one of its goals in increasing the amount of writing students do in school. At the same time, teachers' responses suggest the possibility that this change may have negative ramifications as well, in terms of both instructional impact and test validity. Many teachers maintain that the emphasis on writing is excessive and that instruction has suffered because of the amount of time required for writing. ...[also,] emphasis on writing makes it difficult to judge math competence of some students" (Koretz, Mitchell, Barron, and Keith, 1996, p. xiii).

Rafferty (1993) surveyed urban teachers and staff regarding the MSPAP program. Individuals were asked to respond in Likert fashion to several statements. When the question was "MSPAP will have little effect on classroom practices" 33% agreed or strongly agreed, 24% were uncertain, 42% disagreed or strongly disagreed and 1% did not respond. To the statement "classroom practices are better because of MSPAP" 21% were in agreement, 36% were uncertain, 41% disagreed, and 2% did not answer. To the statement "MSPAP is essentially worthwhile," 24% agreed or strongly agreed, 25% were uncertain, and 48% disagreed or strongly disagreed (3% did not respond). Perhaps a reasonable interpretation of these data is that MSPAP will likely have an impact, but not necessarily a good one.

Koretz, Barron, Mitchell, and Stecher (1996) did a study for Kentucky similar to the one done by Koretz, Mitchell, Barron, and Keith in Maryland. They surveyed the teachers and principals in Kentucky regarding the Kentucky Instructional Results Information System (KIRIS) and found much the same thing as had been found in Maryland. Among other findings,

were the following (Note 3) :

> --90% of the teachers agreed that portfolios made it difficult to cover the regular curriculum (p. 37); --most teachers agreed that imposing rewards and sanctions causes teachers to ignore important aspects of the curriculum (p. 42); --portfolios were cited as having negative effects on instruction almost as often as having had positive effects (p. xi); --almost 90% of the teachers agreed that KIRIS caused them to de- emphasize or neglect untested material (p. xiii); and --other aspects of instruction have suffered as a result of time spent on writing, and emphasis on writing makes it difficult to judge the mathematical competence of some students (p. xv).

McDonnell and Choisser (1997) studied the local implementation of new state assessments in Kentucky and North Carolina. They concluded that

> Instruction by teachers in the study sample is reasonably consistent with the state assessment goals at the level of classroom activities, but not in terms of the conceptual understandings the assessments are measuring. Teachers have added new instructional strategies ... but ... they have not fundamentally changed the depth and sophistication of the content they are teaching. (1997, p. ix.).

Stretcher and Mitchell (1996) reported on the effects of portfolio-driven reform in Vermont and stated that

> The Vermont portfolio assessment program has had substantial positive effects on fourth-grade teachers' perceptions and practices in mathematics. Vermont teachers report that the program has taught them a great deal about mathematical problem solving and that they have changed their instructional practices in important ways (Stretcher and Mitchell, 1996, p. ix).

Smith, Nobel, Heinecke, Seck, Parish, Cabay, Junker, Haag, Tayler, Safran, Penley, and Bradshaw (1997) conducted a study of the consequences of the, now discarded, Arizona Student Assessment Program (ASAP). Although the program had several parts including some norm referenced testing with the Iowa tests, the most visible portion of ASAP was the performance assessment. Teacher opinion of the direction of the effect of ASAP on the curriculum was divided.

> Some defined 'ASAP' as representing an unfortunate and even dangerous de-emphasis of foundational skills, whereas others welcomed the change or saw the new emphasis as encompassing both skills and problem solving. (Smith et al., 1997, p. 40).

Some interesting quotes by teachers found in the Smith et al. report are as follows:

> Nobody cares about basics...The young teachers coming out of college will just perpetuate the problem since they are learning whole language instruction and student-centered classroom. Certainly these concepts

> have their merits, but not at the expense of basics on which education is based (p. 41).
>
> The ASAP ... is designed to do away with 'skills' because kids today don't relate to skills, because they are boring. By pandering to this we are weakening our society, not strengthening it. *It is wrong*! I was told by a state official that teachers would be more like coaches under ASAP. Ask any coach if they teach skills in isolation before they integrate it into their game plan. They will all tell you yes. I rest my case. (Smith et al., 1997, p. 41).

As Smith, et al. report, about two thirds of the teachers believed that "pupils at this school need to master basic skills before they can progress to higher order thinking and problem solving" (1997, p. 41). Forty three percent of the teachers believed that "ASAP takes away from instructional time we should be spending on something more important." (p. 44). In spite of many teachers being unhappy with the content of ASAP, "about 40% of the teachers reported that district scope and sequences had been aligned with ASAP." (p. 46). As the authors report, "changes consequent to ASAP seemed to fall into a typology that we characterized as 'coherent action,' 'compliance only,' 'compromise,' and 'drag.'" (p. 46).

Miller (1998) studied the effect of state mandated performance based assessments on teachers' attitudes and practices in six different contexts (grade level and content areas). Two questions were asked relevant to curricular and instructional impacts.

> "I have made specific efforts to align instruction with the state assessments." (Percents who agreed or strongly agreed ranged from 54.5 to 92.7% across the six contexts.) "I feel that state mandatory assessments have had a negative impact by excessively narrowing the curriculum covered in the classroom." (Percents who agreed or strongly agreed ranged from 28.7 to 46.8%. Only teachers in five of the contexts responded to that question.)

The two questions provide interesting results. While the majority made specific efforts to align instruction, the majority did not feel it resulted in excessively narrowing of the curriculum. However, as Miller pointed out, "the assessments were usually intended to give supplemental information. Consequently, they do not reflect everything that students learn, and only provide a small view of student performance..." (Miller, 1998, pp. 5-6). To align instruction to assessments that provide only a small view of student performance without excessively narrowing the curriculum would seem to be a difficult balancing act.

While more research and opinions could be reviewed, a reasonably summary is that if stakes are high enough and if content is deemed appropriate enough by teachers, there is likely to be a shift in the curriculum and instruction to the content sampled by the test (or the content on the test if the test is not secure). If stakes are low, and/or if teachers believe the assessment is testing developmentally inappropriate materials and/or teaching to the assessment would reduce the amount of time the teachers wish to spend on other -- what they consider more important -- content, the impact is not so obvious.

B. Motivation/morale/stress/ethical behavior of teachers: Increase or Decrease?

Many would argue, quite reasonably, that if we are to improve education, we must depend on the front line educators -- the teachers -- to lead the charge. Do large scale assessments tend to improve the efforts, attitudes and ethical behavior of teachers?

Smith and Rottenberg suggested that external tests negatively affect teachers. As they wrote:

> the chagrin they felt comes from their well-justified belief that audiences external to the school lack interpretive context and attribute low scores to lazy teachers and weak programs (Smith and Rottenberg, 1991, p. 10).

Although I believe they were primarily discussing the effects of traditional assessments, one should expect the same reaction from performance assessments. Audiences external to the school are no more able to infer correct causes of low scores on performance assessments than they are to infer correct causes of low scores on multiple-choice assessments. The inference to lazy teachers and weak programs is equally likely no matter what the test format or test content.

Koretz, Mitchell, Barron and Keith (1996) reported that for the Maryland School Performance Assessment Program (MSPAP):

> Few teachers reported that morale is high, and a majority reported that MSPAP has harmed it. ... 57% of teachers responded that MSPAP has led to a decrease in teacher morale in their school, while only a few (4%) reported that MSPAP has produced an increase (Koretz, Mitchell, Barron and Keith, 1996, p. 24).

Koretz, Barron, Mitchell and Stecher (1996) in the Kentucky study reported that "about 3/4 of teachers reported that teachers' morale has declined as a result of KIRIS" (p. x). Stecklow (1997) reported that there were conflicts in over 40% of Kentucky schools about how to divide up the reward money. So affect was not necessarily high even in the schools that got the rewards! Koretz, Barron, Mitchell and Stecher (1996) also found that principals reported that KIRIS had affected attrition. But the attrition was for both good and poor teachers.

With respect to effort, at least the teachers in Kentucky reported that their efforts to improve instruction and learning had increased (Koretz, Barron, Mitchell and Stecher, 1996, p. 23). But at some point, increased efforts lead to burnout -- and thus attrition increases.

It is commonly believed that some teachers engage in behaviors of questionable ethics when teaching toward, administering and scoring high stakes multiple-choice tests. What about performance assessments? Koretz, Barron, Mitchell and Stecher (1996) reported that in Kentucky

> Appreciable minorities of teachers reported questionable test- administration practices in their schools. About one-third reported that questions are at least occasionally rephrased during testing time, and roughly one in five reported that questions about content are answered during testing, that revisions are recommended during or after testing, or that hints are provided on correct answers. (Koretz, Barron, Mitchell and Stecher, 1996, p. xiii).

In summary, the evidence regarding the effects of large scale assessments on teacher motivation, morale, stress and ethical behavior is sketchy. But what evidence

there is, coupled with what seems logical, suggests that increasing the stakes for teachers will increase efforts, lead to more burnout, decrease morale, and increase the probability of unethical behavior.

C. Motivation and self-concepts of students: Up or down?

With respect to assessment impacts on the affect of students, we are again in a subarea where there is not a great deal of empirical evidence. Logic suggests that the impact on students may be quite different for those tests where the stakes apply to them than for tests where the stakes impact the teachers. Impact surely depends on whose ox is getting gored by the stakes.

Also, the impact should depend on how high the standards are. It is reasonably to believe that the impact of minimal competency tests would be minimal for the large majority of students for whom such tests would not present a challenge. However, for those students who had trouble getting over such a minimal hurdle, the tests probably would both increase motivation and increase frustration and stress -- the exact mix varying on the personality characteristics of the students.

Smith and Rottenberg found that for younger students teachers *believed* that standardized tests

> cause stress, frustration, burnout, fatigue, physical illness, misbehavior and fighting, and psychological distress (Smith and Rottenberg, 1991, p. 10).

That belief of teachers may be true, but certainly does not constitute hard evidence. I come closer to Ebel's view when he suggested that

> Of the many challenges to a child's peace of mind caused by such things as angry parents, playground bullies, bad dogs, shots from the doctor, and things that go bump in the night, standardized tests must surely be among the least fearsome for most children (Ebel, 1976, p. 5).

Lane and Parke (1996), reporting on the consequences of a math performance assessment found that some students developed feelings of inadequacy and, as a result, were less motivated. Miller (1998) found that the percent of teachers responding positively to the statement that performance assessments "increased student confidence" ranged from only 9.1 to 37.6% across five different contexts.

However, Kane et al. (1997) employing a qualitative, case-study methodology and visiting 16 schools ("not confirmed to be representative" --p. xvi) developing and implementing performance assessments reported that

> many interviewees reported that students exhibit a greater motivation to learn and a greater amount of engagement with performance tasks and portfolio assignments than with other types of assignments" (Kane et al., 1997, p. 201).

Koretz, Barron, Mitchell and Stecher (1996) reported that in Kentucky one third of the teachers reported that students' morale had deteriorated and virtually none reported an increase in student morale. They also reported that an emphasis or writing caused students to become tired of writing.

As mentioned earlier, one of the factors effecting student affect is how high the standards are set. Minimum standards are not likely to have a major impact. High standards might. Linn (1994) has pointed out that

> The dual goals of setting performance standards for student certification that are both 'world class' and apply to 'all' students are laudable, but it cannot simply be assumed that only positive effects will result from this press (Linn, 1994, p. 8).

Linn quoted Coffman as follows:

> Holding common standards for all pupils can only encourage a narrowing of educational experiences for most pupils, doom many to failure, and limit the development of many worthy talents (Coffman, 1993, p. 8; quoted from Linn, 1994).

We are simply putting too many students and too many teachers under too much pressure if we hold unrealistically high standards for all students. As Bracey has said, in an article entitled "Variance Happens -- get over it!"

> We are currently in a period that adheres rabidly to an all- children-can-learn philosophy. ... The stance is a philosophical, moral -- almost religious -- posture taken by a wide spectrum of educators and psychologists who ought to know better. ... By telling everyone that all children can learn, we set the stage for the next great round of educational failure when it is revealed that not everyone **has** learned, in spite of our sincere beliefs and improved practices. (Bracey, 1995, p. 22 and 26).

Of course his point is not that some children can not learn anything, but that not everyone can achieve at high standards in academics anymore than everyone can become athletically proficient enough in every sport to play on the varsity teams.

D. True improvement in student learning, or just higher test scores?

> In mandating tests, policy makers have created the illusion that test performance is synonymous with the quality of education (Madaus, 1985, p. 617).

All of us recognize that it is possible for test scores to go up without an increase in student learning on the domain the test supposedly samples. This occurs, for example, when teachers teach the questions on non-secure tests. Teaching too closely to the assessment results in the inferences from the test scores being corrupted. One can no longer make inferences from the test to the domain. The Lake Wobegon effect results. Many of us have written about that (e.g. Mehrens and Kaminski, 1989).

If the assessment questions are secure and the domain the test samples is made public, corrupting reasonable inferences from the scores is more difficult. If the inference from rising test scores of secure tests is that students have learned more of the domain the test samples, that is likely a correct inference. However, those making inferences may not realize how narrow the domain is, or that a test sampling a similar sounding but somewhat differently defined domain might give different results. Of course, if the inference from rising scores is that educational quality has gone up, that may not be true.

1. Improvement on traditional tests

Pipho has reported that:

> Ironically, every state that has initiated a high school graduation test in grade 8 or 9 has reported an initial failure rate of approximately 30%. By 12th grade, using remediation and sometimes twice-a-year retests,

> this failure rate always gets down to well under 5% (Pipho, 1997, p. 673).

Is this *true improvement*, or is it a result of teaching to the test? Recall that these tests are supposedly secure so one cannot teach the specific questions. However, one could limit instruction to the general domain the tests sample. My interpretation is that the increase in scores represents a true improvement on the domain the test samples, but that it does not necessarily follow that it is a true improvement in the students' education.

2. Improvement on performance assessments

What about performance assessments? Do increases in scores indicate necessary improvement in the domain, or an increase in educational quality? Certainly no more so than for multiple-choice assessments, and perhaps less so. Even if specific tasks are "secure," performance assessments are generally thought to be even more "memorable" and reusing such assessments can result in corrupted inferences. If the inference is to only the specific task, there may not be too much corruption, but any inferences to a domain the task represents or to the general quality of education are as likely to be incorrect for performance assessments as for multiple-choice assessments.

Shepard, Flexer, Hiebert, Marion, Mayfield, and Weston (1996) conducted a study investigating the effects of classroom performance assessments on student learning. As they stated:

> Overall, the predominant finding is one of no-difference or no gains in student learning following from the year-long effort to introduce classroom performance assessments. Although we argue subsequently that the small year-to-year gain in mathematics is real and interpretable *based on qualitative analysis, honest discussion* of project effects must acknowledge that *any benefits are small and ephemeral* (Shepard et al., 1996, p. 12, emphasis added).

Others, doing less rigorously controlled studies based on teacher opinion surveys, have been equally cautious in their statements. Khattri et al. (1995), in their study visiting 16 schools stated that

> Only a few teachers said performance-based teaching and assessment helped students learn more and develop a fuller multi-disciplinary understanding (Khattri et al., 1995, p. 82).

Koretz, Barron, Mitchell and Stecher (1996) reported that

> Few teachers expressed confidence that their own schools' increases on KIRIS were largely the results of improved learning (Koretz, Barron, Mitchell and Stecher, 1996, p. xiii)

The authors go on to suggest that

> A variety of the findings reported here point to the possibility of inflated gains on KIRIS--that is, the possibility that scores have increased substantially more than mastery of the domains that the

> assessment is intended to represent (Koretz, Barron, Mitchell and Stecher, 1996, p. xv).

Kane et al. (1997) concluded from their study that

> In the final analysis, the success of assessment reform as a tool to enhance student achievement remains to be rigorously demonstrated (Kane et al., 1997, p. 217).

Miller (1998) asked teachers whether they believed the state mandated performance assessments "have had a positive effect on student learning." Percents across five contexts ranged from 11.3% to 54.7%. When asked whether the tests results were "an accurate reflection of student performance" the percentages ranged from 13.1% to 28.7%.

Finally, for some types of portfolio assessments, one does not even know who did the work. As Gearhart, Herman, Baker, and Whittaker pointed out: "This study raises questions concerning validity of inferences about student competence based on portfolio work." (Gearhart et al., 1993, p. 1).

3. Conclusions about increases

In conclusion, there is considerable evidence that students' pass rates increase on secure high-stakes (mostly multiple-choice) graduation tests. There is at least some reason to believe that students have increased their achievement levels on the specific domains the secure tests are measuring. (Of course, if supposedly secure tests are not actually secure the inference that increased scores indicate increased achievement could be incorrect.) There is less evidence about increases in scores for performance assessments. While it is true that some states (e.g. Kentucky) have shown remarkable gains in scores, evidence points to the possibility that the gains are inflated and there is generally less confidence that achievement in the represented domain has also increased. In neither case can we necessarily infer that quality of education has increased. That inference cannot flow directly from the data. Rather, it must be based on a philosophy of education that says an increase in the domain tested represents an increase in the quality of education. As Madaus stated, it is an illusion to believe at an abstract level that test performance is *synonymous with* quality of education. Nevertheless, test performance can *inform us about* the quality of education -- at least about the quality of education on the domain being assessed.

E. Restore public confidence or provide data for critics?

At an abstract level, it seems philosophically wrong and politically shortsighted for educators to argue against the gathering of student achievement data for accounting and accountability purposes. My own belief is that an earlier stance of the NEA against standardized tests resulted in the public wondering just what it was the educators were trying to hide. I suspect the NEA stance contributed to the action in the state departments that Womer mentioned in 1984. Certainly the public has a right to know something about the quality of the schools they pay for and the level of achievement their children are reaching in those schools.

Some educators strongly believe -- with some supporting evidence -- that the press has incorrectly maligned the public schools (e.g. Bracey, 1996, and Berliner and Biddle, 1995). While their views have not gone unchallenged (see Stedman, 1996) it does seem true that bad news about education travels faster than good news about education. Will the data from large scale assessments change the public's views?

The answer to the above question depends, in part, upon whether the scores go up, go down, or stay the same. It also depends upon whether the public thinks we are measuring anything worthwhile. And it also depends upon how successful we are at communicating the data and communicating what reasonable inferences can be drawn from the data.

Scores generally have been going up in Kentucky, but it has not resulted in all the press highlighting the great job educators are doing in Kentucky. For example Stecklow (1997), in writing about the Kentucky approach suggests:

> It has spawned lawsuits, infighting between teachers and staff, anger among parents, widespread grade inflation -- and numerous instances of cheating by teachers to boost student scores. (Stecklow, 1997, p. 1).

A conclusion of the evaluation done of KIRIS by The Evaluation Center of Western Michigan University stated that

> ...**all** the cited evidence suggests stakeholders have questions concerning the legitimacy, validity, reliability, and fairness of the KIRIS assessment. We have no evidence to suggest that parents think the assessment component of KIRIS is a fair, reliable, and valid system (The Evaluation Center, 1995, p. 20).

The Stecklow quote is not a ringing endorsement of the program or the quality of education in Kentucky. The Evaluation Center quote is not a ringing endorsement of the quality of the data in the assessment. But, in general, the public is happier with high scores than they are with low scores -- often considering which district to live in based on published test scores. (The public may be making two quite different inferences from these scores. One, probably an incorrect inference, is that the district with the higher scores has better teachers or a better curriculum. A second, correct inference is that if their children attend the district with the higher scores they will be more likely to be in classes with a higher proportion of academically able fellow students.)

There is currently considerable concern about whether the newer "reform" assessments cover the correct content. Reform educators were not happy with minimum competency tests covering basics; but the public is not happy with what they perceive to be a departure from teaching and testing the basics. Baker, Linn, and Herman (1996) talk about the crisis of credibility that performance assessments suffer based on a large gap between the views of educational reformers and segments of the public. McDonnell (1997) stated that

> ...the political dimensions of assessment policy are typically overlooked. Yet because of their link to state curriculum standards, these assessments often embody unresolved value conflicts about what content should be taught and tested, and who should define that content. (McDonnell, 1997, p. v).

As McDonnell pointed out, there are fundamental differences between what educational reformers and large segments of the public believe should be in the curriculum.

> The available opinion data strongly suggest that the larger public is skeptical of new

> curricular approaches in reading, writing, and mathematics (McDonnell, 1997, p. 67).

The truth of this can be seen by the fight over the mathematics standards in California.

The press and public seem either reasonably unimpressed by the data educators provide and/or make incorrect inferences from it. What can change that? We need to gather high quality data over important content and communicate the data to the public in ways that encourage correct inferences about students' levels of achievement. We need to be especially careful to discourage the public from making causative inferences if they are not supported by the research data.

IV. ARE THE CONSEQUENCES GOOD OR BAD?

It should be obvious by now that I do not believe we have a sufficient quantity of research on the consequences of assessment. Further, the evidence we do have is certainly not of the type from which we can draw causative inferences, which seems to be what the public wants to do. Given the evidence we do have, can we decide if it suggests the consequences of assessment are positive or negative? If the evidence were better, could we decide if the consequences are positive or negative? I maintain that each of us can decide, but we may well disagree. Interpreting the consequences as being good or bad is related to differences in convictions about the proper goals of education. Let us look at the evidence regarding each of the five potential benefits and/or dangers with respect to the quality of the consequences.

A. Curricular and Instructional Reform

While there is no proven cause and effect relationship between assessment and the curriculum content or instructional strategies there is some evidence and compelling logic to suggest that high stakes assessments can influence both curriculum and instruction. Is this good or bad? It is a matter of one's goals. Reform educators were dismayed to think that minimum competency tests using multiple-choice questions were influencing curriculum and instruction. They pushed for performance assessments, not because they abhorred tests influencing curriculum and instruction, but because they wanted the tests to have a different influence.

The public was not dismayed that educators tested the basics -- they rather approved. They believe (some evidence suggests incorrectly) that educators have moved away from basics and are dismayed. Obviously the narrowing and refocusing of the curriculum and instructional strategies are viewed as either negative or positive depending upon whether the narrowing and refocusing are perceived to be toward important content. Educators and the public do not necessarily agree about this.

B. Increasing Teacher Motivation and Stress

Increasing teacher stress may be perceived as good or bad -- depending on whether one believes teachers are lazy and need to be slapped into shape or whether one believes (as I do) that teachers already suffer from too much job stress.

C. Changing Students' Motivation or Self Concept

We might all favor an increase in student motivation. I, for one, do not believe a

major problem in education in the United States is that students are trying too hard to learn too much. But some educators do worry about the stress that tests cause students (recall the quote from Smith and Rottenberg). There is such a thing as "test anxiety" (more accurately called evaluation anxiety), but many would argue that occasional state anxiety is a useful experience -- perhaps helping individuals to learn how to cope with anxiety and to treat stress as eustress rather than distress.

But what if assessment lowers students' self concepts? Again, this could be either good or bad -- depending on whether one believes students should have a realistic view of how inadequate their knowledge and skills are. (Recall that in Japan, whose students outperform U.S. students, the students do not feel as confident in their math competencies as do U.S. students.) As one colleague has pointed out to me, we are not necessarily doing students a favor by allowing them to perceive themselves as competent in a subject matter if that, indeed, is not the truth (Ryan, personal communication, 1997).

D. Increased Scores on Assessments

Surely this is good -- right? It again depends. It depends on whether the gains reflect improvement on the total domain being assessed or just increases in scores, whether we care about the tested domain, and whether, as a result of the more focused instruction, other important domains (not being tested) suffer.

E. Public Awareness of Student Achievement

Is public awareness of how students score on assessments good or bad? Obviously one answer is that it depends on whether valid inferences are drawn from the data. One part of the validity issue is whether the scores truly represent what students know and can do. Another part of the validity issue is whether the public draws causative inferences that are not supported by the data.

In addition to the question of whether the inferences are valid, there is the issue of how the public responds. Would negative news stimulate increased efforts by the public to assist educators --e.g. by trying to ensure children start school ready to learn, by providing better facilities, by insisting their children respect the teachers? Or would negative news result in more rhetoric regarding how bad public schools are, how bad the teachers are, and how we should give up on them and increase funding to private schools at the expense of funding public schools? Would positive news result in teachers receiving public accolades and more respect or would the public then place public education on a back burner -- because the "crisis" was over?

While I come down on the side of giving the public data about student achievements, the communication with the public must be done with great care. I believe there is a propensity for the public (at least the press) to engage in inappropriate blaming of educators when student achievement is not as high as desired. I am reminded of Browder's (1971) suggestion that accountability boils down to who gets hanged when things go wrong and who does the hanging. Educators have good reason to believe that they are the ones who will get hanged and the public, abetted by the press, will do the hanging.

Dorn has stated that "test results have become the dominant way states, politicians, and newspapers describe the performance of schools." (1998, p. 2). He was not suggesting that was a positive happening.

V. WHAT VARIABLES CHANGE PROBABILITIES FOR GOOD OR BAD IMPACTS

Since *whether* consequences are good or bad is partly a matter of one's educational values, it is difficult to answer this question. Nevertheless, I will provide a few comments.

A. Impact should be (and likely is) related to purposes.

As has been mentioned, there are two major purposes of large scale assessments: to drive reform, and to see if reform practices have had an impact on student learning. These are somewhat contradictory purposes, because current reformers believe assessment should be "authentic" if it is to drive reform and most authentic assessment is not very good measurement -- at least by any conventional measurement criteria.

B. Impact (and purposes) are likely related to test content, and the *public involvement* in determining content and content standards.

> Successful assessment reform needs to be an open and inclusive process, supported by a broad range of policy makers, educators, and the public, and closely tied to standards in which parents and the community have confidence. (The CRESST Line, 1997, p. 6).

The impact is not likely to be positive in any overall sense if the public has not bought into the content standards that are being assessed.

One can also expect some problems if the content and test standards are set too high. The politically correct rhetoric that "all children can learn to high levels" has yet to be demonstrated as correct. Recall the quote by Coffman given earlier. Recall also Bracey's article entitled: Variance happens--get over it. Or, as a colleague once said, would we require the PE instructor to get all students up to a level where they are playing on the varsity team?

C. Impact may be related to item or test format.

If the issue is whether the overall impact is good or bad, there is not much evidence that item or test format matters. The abstract notion that teaching to improve performance assessment results means educators will be teaching like they should be teaching whereas teaching to improve multiple-choice test scores means teaching is of poor quality is just nonsense.

D. Impact may be related to the quality of the assessment (perceived or real) and the assessment procedures (e.g. test security and reporting practices).

If educators do not believe the assessments provide high quality data, they may not pay much attention to them. Cunningham made this point very forcefully in discussing the Kentucky Instructional Results Information System (KIRIS) program:

> As teachers begin to realize that the test has no legitimacy and that it is too technically deficient to be influenced by how they teach, they will stop paying attention to it. ... Measurement driven instruction does not work when teachers fail to see the connection between measurement and instruction. (Cunningham, no date, p. 2).

Whatever one believes about the technical adequacy of KIRIS, Cunningham's general point would seem accurate: If teachers do not see any connection between the assessment results and their instructional approaches the measurement is unlikely to impact instruction.

Another example comes from a paper Smith et al. (1997) wrote on the consequences of the Arizona Student Assessment Program (ASAP). Some teachers believed that the ASAP skills were not developmentally appropriate (p. 38), some objected to what they perceived as poor-quality rubrics and to the subjectivity of the scoring process (p. 39), some thought ASAP was just a fad and one teacher referred to ASAP as Another Stupid Aggravating Program (p. 43). Again, the point is not whether the teachers' perceptions were correct. But if they perceive the assessment quality to be poor, they are not likely to be very positively impacted by it.

E. Impact may depend upon degree of sanctions.

Some limited research evidence regarding this variable comes from a McDonnell and Choisser (1997) study. They investigated the local implementation of state assessments in North Carolina and Kentucky. As they suggested, Kentucky's program involved high stakes for schools and educators, with major consequences attached to the test results. The North Carolina assessment had no tangible consequences attached to it. However

> teachers in the two state samples perceive the new assessments in much the same way and take them equally seriously. With few exceptions, their teaching reflects the assessment policy goals of their respective states to a similar degree. (McDonnell and Choisser, 1997, p. ix).

Of course, the North Carolina assessments have some consequences. Results are presented in district 'report cards' and in school building improvement results. And, at the time the study was conducted, McDonnell and Choisser reported that

> probably the most potent leverage the assessment system has over the behavior of teachers is the widespread perception that local newspapers plan to report test scores not just by individual school, which has been done traditionally, but also by specific grade level and even by classroom (1997, p. 16).

One can imagine why teachers in North Carolina might think the stakes were fairly high in spite of no state financial rewards or sanctions.

Thus, in spite of the evidence which shows no distinctions between Kentucky, which used financial rewards, and North Carolina, which simply made the scores available, the perceived stakes to the teachers may not have been much different. I continue to believe that as stakes increase, dissatisfaction increases, fear increases, cheating increases, and lawsuits increase. However, efforts may also increase to improve scores and, if the procedures are set up to make it difficult to improve scores without improving competence on the domain, student learning should increase also.

F. Impact may relate to level of professional development.

Unfortunately, many current reform policies concentrate more on standards and assessments than they do the professional development of teachers. In the Arizona SAP program for example, only 19% of the teacher felt that adequate professional development had been provided (Smith et al., 1997). Combs, in his critique of top-down reform mandates stated that: "Things don't change people; people change things." (quoted from Smith et al., 1997, p.50). As Smith et al. pointed out in their review, Cohen (1995, p. 13) had noted the apparent anomaly in the systemic reform movement and accountability intentions. Motivated by perceptions that public schools are failing,

> advocates of systemic reform propose to radically change instruction, and for that they must rely on teachers and administrators. But these agents of change are the very professionals whose work reformers find so inadequate. (Quoted from Smith et al., 1997, p. 105).

VI. CONCLUSIONS

So, what can we conclude about the consequences of assessment? I list a dozen.

A. Purposes and Expectations

1. There are a variety of purposes for and expectations regarding the consequences of assessment. Some of these may be unrealistic. "Evaluation and testing have become *the* engine for implementing educational policy" (Petrie, 1987, p. 175).

B. Need for Evidence

2. Scholars seem to agree that it is unwise, illogical, and unscholarly to just assume that assessments will have positive consequences. There is the potential for both positive consequences and negative consequences.

C. Quantity and Quality of the Evidence

3. It would profit us to have more research.

4. The evidence we do have is inadequate with respect to drawing any cause/effect conclusions about the consequences. If instruction changes concomitant with changes in both state curricular guidelines and state assessments, how much of the change was due to which variable?

D. Evaluating the Evidence

5. Not everyone will view changes (e.g. reforming curriculum in a particular way) with the same affect. Some will think the changes represent positive consequences and others will think

the changes constitute negative consequences.

E. Curricular and Instructional Consequences

6. High stakes assessments probably do impact both curriculum and instruction, but assessments alone are not likely as effective as they would be if there were more teacher professional development.

7. Attempts to reform curriculum in ways neither front line teachers nor the public support seems unwise.

F. Impact on Teachers

8. High stakes assessments increase teacher stress and lower teacher morale. This seems unfortunate to me, but may make others happy.

9. Assessments can assist both students and teachers in evaluating whether the students are achieving at a sufficiently high level. This seems like useful knowledge.

G. Impact on Test Scores and Student Learning

10. High stakes assessments will result in higher test scores. Both test security and the opportunity to misadminister or mis-score tests must be considered in evaluating whether higher scores represent increased knowledge. If the test items are secure (and reused items are not memorable), and if tests are administered and scored correctly, it seems reasonably to infer that higher scores indicate increased achievement in the particular domain the assessment covers. That is good if the domain represents important content and if teaching to that domain does not result in ignoring other equally important domains. If tests are not secure, or are incorrectly administered or scored, there is no reason to believe that higher scores represent increased learning.

H. Impact on Public

11. The public and the press are more likely to use what they believe to be "inadequate" assessment results to blame educators than to use "good" results to praise them. They will continue to make inappropriate causative inferences from the data. The public will not be impressed by assessments over reform curricula they consider irrelevant.

I. Confounding Format, Content and Stakes in Considering Consequences

12. There has been a great deal of confounding of item format, test content and the stakes. Which format is used probably makes far less difference than how it is used.

References

Airasian, P. W. (1988). Measurement driven instruction: A closer look. *Educational Measurement: Issues and Practice*, 7(4), 6-11.

Anderson, B. L. (1985). State testing and the educational community: Friends or foes? *Educational Measurement: Issues and Practice,* 4(2), 22-25.

Anderson, B., and Pipho, C. (1984). State-mandated testing and the fate of local control. *Phi Delta Kappan,* 66, 209-212.

Baker, E. L., Linn, R. L., and Herman, J. L. (1996, Summer). CRESST: A continuing mission to improve educational assessment. *Evaluation Comment*.

Baker, E.L., O'Neil, H.F., and Linn, R.L. (1993). Policy and validity prospects for performance-based assessment. *American Psychologist*, 48(12), 1210-1218.

Berliner, D., and Biddle, B. (1995). *The manufactured crisis: Myths, fraud, and the attack on America's public schools.* New York: Addison-Wesley.

Bracey, G.W. (Fall, 1995). Variance happens -- get over it! *Technos*, 4(3), 22-29.

Bracey, G.W. (1996). International comparisons and the condition of American education, *Educational Researcher*, 25(1), 5-11.

Browder, L.H. Jr. (1971). *Emerging patterns of administrative accountability*. Berkeley, CA: McCutchan.

Chudowsky, N., and Behuniak, P. (1997, March). Establishing consequential validity for large-scale performance assessments. Paper presented at annual meeting of the National Council of Measurement in Education, Chicago, IL.

CRESST. (1997, Spring). Analyzing statewide assessment reforms. *The CRESST Line*.

Cunningham, G. K. (No date). Response to the response to the OEA panel report. University of Louisville.

Dorn, S. (1998). The political legacy of school accountability systems. *Education Policy Analysis Archives*, 6, No. 1 (Entire Issue). (Available online at http://epaa.asu.edu/epaa/v6n1.html).

Ebel, R.L. (1976). The paradox of educational testing. *Measurement in Education*, 7(4), 1-12.

The Evaluation Center. (1995). *An independent evaluation of the Kentucky Instructional Results Information System (KIRIS)* [Report conducted for The Kentucky Institute for Education Research]. Western Michigan University.

Floden, R.E. (1998). Personal communication.

Froomkin, D. (Sept. 29, 1997). National education tests: An introduction. *Back to the top* [on line], Digital Ink Company.

Gearhart, M., Herman, J.L., Baker, E.L., and Whittaker, A.K. (July 1993). *Whose work is it? A question for the validity of large-scale portfolio assessments.* CSE Technical Report 363. Center for the study of evaluation, National Center for Research on Evaluation Standards, and Student Teaching, Graduate School of Education, University of California, Los Angeles.

Green, D. R. (1997, March). *Consequential aspects of achievement tests: A publisher's point of view*. Paper presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, Chicago, IL.

Haney, W., and Madaus, G. (1989). Searching for alternatives to standardized tests: Whys, whats, and whithers. *Phi Delta Kappan*, 70(9), 683-687.

Herman, J.L., Klein, D.C.D., Heath, T.M. and Wakai, S.T. (December, 1994). *A first look: Are claims for Alternative assessment holding up?* CSE Technical Report 391. National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles.

Jones, L.V. (1997). *National tests and education reform: Are they compatible?* William H. Angoff Memorial Lecture Series. Educational Testing Service.

Kane, M. B., Khattri, N., Reeve, A. L., and Adamson, R. J. (1997). *Assessment of student performance*. Washington D.C.: Studies of Education Reform, Office of Educational Research and Improvement, U.S. Department of Education.

Khattri, N., Kane, M. B., and Reeve, A. L. (1995). *How performance assessments affect teaching and learning* [Research Report]. Educational Leadership.

Koretz, D. (1996). Using student assessments for educational accountability. In E. A. Hanushek and D.W. Jorgenson (Eds). *Improving America's schools: The role of incentives*. (pp. 171-195). National Academy Press, Washington, DC.

Koretz, D, Barron, S., Mitchell, K., and Stecher, B. (1996, May). *Perceived effects of the Kentucky Instructional Results Information System (KIRIS)* . Institute on Education and Training, RAND.

Koretz, D., Mitchell, K., Barron, S., and Keith, S. (1996). *Final report: Perceived effects of the Maryland school performance assessment program* [CSE Technical Report 409]. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST) (57 pages).

Kuhs, T., Porter, A., Floden, R., Freeman, D., Schmidt, W., and Schwille, J. (1985). Differences among teachers in their use of curriculum-embedded tests. *The Elementary School Journa*l, 86(2), 141-153.

Lane, S. (1997, March). Framework for evaluating the consequences of an assessment

program. Paper presented at the annual meeting of the National Council of Measurement in Education, Chicago, IL.

Lane, S., and Parke, C. (1996, April). Consequences of a mathematics performance assessment and the relationship between the consequences and student learning. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Langdon, C.A. (1997). The fourth Phi Delta Kappan poll of teachers' attitudes toward the public schools. *Phi Delta Kappan*, 79(3), 212- 220.

Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis*, 15(1), 1-16.

Linn, R.L. (1994). Performance Assessment: Policy promises and technical measurement standards. *Educational Researcher*, 23(9), 4-14.

Linn, R.L., Baker, E.L., and Dunbar, S.B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, 20(8), 15-21.

Linn, R.L., and Herman, J.L. (1997, February). *Standards-led assessment: Technical and policy issues in measuring school and student progress.* CSE Technical Report 426. National Center for Research on Evaluation, Standards, and Student Testing (CRESST) Center for the Study of Evaluation, Graduate School of Education and Information Studies, University of California, Los Angeles.

Madaus, G. F. (1985). Test scores as administrative mechanisms in educational policy. *Phi Delta Kappan*, 66(9), 611-617.

Madaus, G.F. (1991). The effects of important tests on students: Implications for a National Examination System. *Phi Delta Kappan*, 73(3), 226-231.

Madaus, G.F., West, M.M., Harmon, M.C., Lomax, R.G. and Viator, K.A. (1992, October). The influence of testing on teaching math and science in grades 4-12. Executive Summary. National Science Foundation Study, Center for the Study of Testing, Evaluation, and Educational Policy, Boston College, Chestnut Hill, Massachusetts.

Mayes, M. (1997, August 30). Test results make school chief smile. "The Lansing State Journal," pp. 1A, 5A.

McDonnell, L. M. (1997). The politics of state testing: Implementing new student assessments [CSE Technical Report 424]. University of California, Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

McDonnell, L.M. and Choisser, C. (1997, September). Testing and teaching: Local implementation of new state assessments. CSE Technical Report 442. National Center for Research on Evaluation, Standards, and Student Testing (CRESST) Center for the Study of Evaluation (CSE) Graduate School of Education and Information Studies, University of California, Los Angeles, CA.

Mehrens, W.A. and Kaminski, J. (1989). Methods for improving standardized test

scores: Fruitful, fruitless or fraudulent? *Educational Measurement: Issues and Practices*, 8(1), 14-22.

Miller, M.D. (1998, February). Teacher uses and perceptions of the impact of statewide performance-based assessments. Council on Chief State School Officers. State Education Assessment Center, Washington, D.C.

Petrie, H.G. (1987). Introduction to "evaluation and testing." *Educational Policy*, 1, 175-180.

Pipho, C. (1997). Standards, assessment, accountability: The tangled triumvirate. *Phi Delta Kappan*, 78(9), 673-674.

Pomplun, M. (1997). State assessment and instructional change: A path model analysis. *Applied Measurement in Education*, 10(3), 217- 234.

Porter, A. C., Floden, R. E., Freeman, D. J., Schmidt, W. H., and Schwille, J. P. (1986). Content determinants [Research Series No. 179]. Michigan State University, East Lansing, MI: Institute for Research on Teaching.

Rafferty, E. A. (1993, April). Urban teachers rate Maryland's new performance assessments. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Reckase, M. D. (1997, March). Consequential validity from the test developers' perspective. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Roeber, E., Bond, L.A., and Braskamp, D. (1997). Trends in statewide student assessment programs, 1997. North Central Regional Educational Laboratory and the Council of Chief State School Officers.

Rose, L.C., Gallup, A.M. and Elam, S.M. (1997). The 29th annual Phi Delta Kappa/Gallup poll of the public's attitudes toward the public schools. *Phi Delta Kappan*, 79(1), 41-58.

Ryan, J.M. (1997). Personal communication.

Seidman, R. H. (1996, July 24). National education 'Goals 2000': Some disastrous unintended consequences. *Education Policy Analysis Archives*, 4, No. 11 (Entire Issue). (Available online at http://epaa.asu.edu/epaa/v4n11/).

SERVE. (1994). A new framework for state accountability systems [Special report of The Southeastern Regional Vision for Education].

Shepard, L. A., (1991). Will national tests improve student learning? *Phi Delta Kappan*, 72, 232-238.

Shepard, L. A., Flexer, R. J., Hiebert, E. H., Marion, S. F., Mayfield, V., and Weston, T. J. (1996, Fall). Effects of introducing classroom performance assessments on student learning. *Educational Measurement: Issues and Practices*, pp. 7-18.

Smith, M. L., Noble, A., Heinecke, W., Seck, M., Parish, C., Cabay, M., Junker, S., Haag, S., Tayler, K., Safran, Y., Penley, Y., and Bradshaw, A. (1997). Reforming schools by reforming assessment: Consequences of the Arizona student assessment program (ASAP): Equity and teacher capacity building [CSE Technical Report 425]. University of California, Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Smith, M.L., and Rottenberg, C. (1991). Unintended consequences of external testing in elementary schools. *Educational Measurement: Issues and Practice*, 10(4), 7-11.

Stecher, B.M. and Mitchell, K.J. (1995, April). Portfolio-driven reform: Vermont teachers' understanding of mathematical problem solving and related changes in classroom practice. CSE Technical report 400. National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Graduate School of Education and Information Studies, University of California, Los Angeles.

Stecklow, S. (1997, September 2). Kentucky's teachers get bonuses, but some are caught cheating. "The Wall Street Journal," pp. A1 and A5.

Stedman, L.C. (1996, January 23). The achievement crisis is real: A review of *The manufactured crisis. Education Policy Analysis Archives*, 4, No. 1 (Entire issue). (Available online at http://epaa.asu.edu/epaa/v4n1.html).

Taleporos, E. (1997, March). Consequential validity. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Womer, F.B. (1984). Where's the action? *Educational Measurement: Issues and Practice*, 3(3), 3.

Notes

1. This paper is a slight revision of the 1998 Vice Presidential address for Division D, American Educational Research Association, presented in San Diego. I would like to thank Bob Floden and Joe Ryan for helpful comments on a previous draft of this paper. Opinions expressed are those of the author and not necessarily those of the two reviewers.
2. As Jones has wondered "Can he really mean that?" (Jones, 1997, p. 3).
3. Space does not permit me to do justice to this very thorough report. I urge readers to obtain the report and study it carefully. Most of these state assessments equate through anchor items, and these items may not be totally secure.

About the Author

William A. Mehrens

Professor of Measurement
462 Erickson Hall
Michigan State University
East Lansing, MI 48824

517-355-9567
FAX 517-353-6393

WMehrens@pilot.msu.edu

WILLIAM A. MEHRENS is a professor of measurement at Michigan State University. He received his Ph.D. in educational psychology from the University of Minnesota in 1965. His interests include educational testing in general, legal issues in high-stakes testing, teaching to the test, and performance assessment. He has been elected to office in several professional organizations including the presidency of both the National Council on Measurement in Education (NCME) and the Association for Measurement and Evaluation in Guidance (currently called the Association for Assessment in Counseling). He is the immediate past Vice President of Division D of the American Educational Research Association (AERA). He is the author or co-author of several major textbooks and many articles. Honors include the NCME Award for Career Contributions to Educational Measurement, 1997; a University of Nebraska-Lincoln Teachers College Alumni Association Award of Excellence, 1997; an AACD Professional Development Award, 1991; MSU Distinguished Faculty Award, 1983; APA Division 15 Fellow, 1984; APA Division 5 Fellow, 1978; and Pi Mu Epsilon, National honorary mathematics fraternity, 1958.

The World Wide Web address for the *Education Policy Analysis Archives* is http://olam.ed.asu.edu/epaa

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. (602-965-2692). The Book Review Editor is Walter E. Shepherd: shepherd@asu.edu . The Commentary Editor is Casey D. Cobb: casey@olam.ed.asu.edu .

EPAA Editorial Board