DOCUMENT RESUME

ED 454 251 TM 032 854

AUTHOR Kane, Michael
TITLE The Role of Policy Assumptions in Validating High-stakes Testing Programs.
PUB DATE 2001-04-00
NOTE 36p.; Paper presented at the Annual Meeting of the American Educational Research Association (Seattle, WA, April 10-14, 2001).
PUB TYPE Opinion Papers (120) -- Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Educational Policy; Elementary Secondary Education; *Graduation Requirements; *High Stakes Tests; State Programs; Test Use; *Testing Programs; *Validity
IDENTIFIERS Stakeholders

ABSTRACT

L. Cronbach has made the point that for validity arguments to be convincing to diverse audiences, they need to be based on assumptions that are credible to these audiences. The interpretations and uses of high stakes test scores rely on a number of policy assumptions about what should be taught in schools, and more specifically, about the content standards and performance standards that should be applied to students and schools. For example, a high school graduation test can be developed as a test of minimal competence for the world of work or as a measure of proficiency in the skills needed in college. The assumptions built into the assessment need to be subjected to scrutiny and criticism if a strong case is to be made for the validity of the proposed interpretation. Stakeholder views are a critical part of the evaluation of the policy assumptions implicit in any testing program. The point is made that much of the current practice in the validation of high stakes testing programs, including high school graduation tests, is seriously flawed because only a part of the interpretive argument is evaluated. (Contains 1 table and 42 references.) (SLD)

The Role of Policy Assumptions in Validating High-stakes Testing Programs

Michael Kane
UW, Madison

Paper presented at the annual meeting of the American Educational Research Association, Seattle, 2001

The Role of Policy Assumptions in Validating High-stakes Testing Programs

Abstract

Cronbach has made the point that for validity arguments to be convincing to diverse audiences, they need to be based on assumptions that are credible to these audiences. The interpretations and uses of high-stakes test scores rely on a number of policy assumptions about what should be taught in schools, and more specifically, about the content standards and performance standards that should be applied to students and schools. For example, a H.S. graduation test can be developed as a test of minimal competence for the world of work, or as a measure of proficiency in the skills needed in college. The assumptions built into the assessment need to be subjected to scrutiny and criticism, if a strong case is to be made for the validity of the proposed interpretation, and stakeholder views are a critical part of the evaluation of the policy assumptions implicit in any testing program.

Tests are routinely used to make educational decisions. When these decisions have potentially serious consequences, the testing program is said to be "high-stakes." Note that it is their consequences that insert the "high stakes."

It is also the potential consequences that generate the controversy surrounding these tests, with the proponents (Ward, 2000) looking forward to a wide range of positive consequences, and the critics (Haney, 1991; Eisner, 2001; Thompson, 2001) fearing a rash of negative consequences. There does not seem to be much disagreement about the legitimacy of using the actual or anticipated consequences of the testing program in arguments for or against such testing programs.

Validation: Interpretive Arguments and Validity Arguments

According to the Standards for Educational and Psychological Testing (AERA, APA, NCME, 1999, p. 9), "Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests." The test itself is not validated, and test scores *per se* are not validated. It is the interpretation determined by the proposed use that is validated (Cronbach, 1971; Messick, 1989).

A validity argument, provides an overall evaluation of the plausibility of the proposed interpretation and uses of test scores (House, 1980; Cronbach, 1988). It aims for a cogent presentation of all of the evidence relevant to the proposed interpretation, and to the extent possible, the evidence relevant to plausible alternate interpretations.

In order to evaluate the plausibility of a test-score interpretation (i.e., in order to develop a validity argument), it is necessary to be clear about what the interpretation claims. A proposed interpretation can be defined in some detail by specifying it as an

interpretive argument, or network of inferences and supporting assumptions leading from scores to conclusions and decisions (Kane, 1992; Shepard, 1993; Crooks, Kane & Cohen, 1996). The interpretive argument serves two functions. It provides an explicit statement of the proposed interpretation/use, and it provides a framework for developing a validity argument.

Given a clear statement of the proposed interpretation in terms of an explicitly stated interpretative argument, the evidence needed to validate the interpretation is the evidence needed to evaluate the inferences and assumptions in the interpretive argument. Interpretive arguments generally contain a number of inferences and assumptions, and the validation effort should presumably focus on studies that are most relevant to the inferences and assumptions in the interpretive argument, preferably paying particular attention to the weakest links in the argument.

For example, the scores on an achievement test consisting of mathematics problems could be interpreted as a measure of problem-solving skill in mathematics. The first step in the interpretation would involve an evaluation of the student's performance on each problem, resulting in the assignment of a score to the student's solution to the problem. The score on all of the problems in the test could then be combined into a single observed score for the student. The observed score could be generalized from the performances actually observed to a universe of possible performances on similar problems under similar circumstances. The scores could then be extrapolated to conclusions about expected performance on the domain of mathematics problems associated with "problem-solving in mathematics." For the interpretive argument as a whole to be considered valid, each of the inferences in the argument should be supported by appropriate evidence.

Validation involves a clear statement of the proposed interpretation, by explicitly stating what it claims and what it assumes, the identification of potential competing interpretations, and the development of evidence supporting the inferences and assumptions in the interpretation and refuting competing interpretations (Cronbach, 1988, Messick, 1989). The validity argument evaluates the plausibility of the interpretive argument by examining whether the conclusions follow from the assumptions and whether the assumptions are reasonable, a priori, or are supported by data.

Note that test-score interpretations are constructs in the sense that they are constructed by some person or persons. Furthermore, a test score may have several legitimate interpretations and may be used to make different kinds of decisions, each with its own interpretive argument. In the example introduced earlier, a score on the mathematics test might be interpreted simply in terms of skill in solving the kind of problem included in the test. This basic interpretation involves the evaluation of performance and the generalization of the observed performance to a larger universe of "similar" performances, but it does not require an extrapolation to any other kind of performance. It is simply an inference to expected performance on a particular kind of task, based on a sample of that kind of performance. This simple interpretation is relatively easy to support using content-related evidence (Cronbach, 1971; Messick, 1989) and evidence for the generalizability of the scores (Brennan, 1992; Cronbach, Gleser, Nanda, & Rajaratman, 1972).

This interpretation could be extended to conclusions about expected performances on other kinds of tasks in the same context, to the same kind of tasks in other contexts, or to different tasks in different contexts. These additional inferences

require additional evidence for their support, in particular, evidence that performance on one kind of task or in one context is related to performance on the other kinds of tasks in other contexts.

A deeper and more sophisticated interpretation sees the scores on the test as indications of level of achievement on certain procedural and problem-solving skills postulated by a model of performance. The model-based, theoretical interpretation is considerably richer than a simple generalization from performance on a sample of tasks to expected performance on the universe of tasks from which the sample was drawn, and as a result, requires more evidence for its support. In particular, the validity argument for model-based, theoretical interpretations will require evidence for the validity of the theory of performance as well as evidence that the assessments can be interpreted in terms of the theory. An interpretation of test scores in terms of a theoretical model depends on evidence for the model and for the relationship between the observed scores and terms in the model.

The test scores could also be used to make a wide variety of decisions. They could be used to make decisions about end-of-course grades, about graduation from high school, about college admission, or about being hired. It is very unlikely that any single test would serve all of these decisions equally well, and the appropriateness of the test for any of these decisions would be open to question. The use of the test scores to make a particular kind of decision requires evidence supporting the appropriateness of the decision.

In general, the evidence required for validation depends on the proposed interpretation, and it is entirely possible for one or more of these interpretations to be valid, while other of these interpretations are invalid. For example, it is possible that the

test scores provide a good indication of an examinee's skill in solving the kind of problem included in the test, but provide a very poor indication of skill in any wider set of problems or in any other context. On a deeper level, it may be that the test provides a valid measure of skill in solving some kind of problem, but does not provide a good indication of the student's mastery of particular procedural and problem-solving skills. The test can remain the same, and the population of examinees can remain the same, and the scores can remain the same, and yet, validity can vary from one interpretation to another.

I risk belaboring this point, because it is critical to my thesis, and because it is too often passed over lightly as a proposed interpretation is taken for granted and accepted without criticism. Of particular importance in what follows is the possibility that the assumptions supporting a descriptive interpretation of test scores in terms of expected performance in some domain (e.g., the content-domain specified by Test Standards) can be well supported by evidence, while the additional assumptions needed to support a particular use of these test scores are not verified, or even worse, are contradicted by available data.

Descriptive Interpretations and Decision-based Interpretations

As noted above, it is the interpretation of test scores that is validated. The test scores can be interpreted in terms of a specific attribute (e.g., achievement in specified content domain) without specifying any particular use of the scores, or it can be directed toward a specific use of the test scores. I will refer to interpretations that estimate some variable for the examinees being tested, without specifying any particular uses for the test scores, as descriptive interpretations.

I will refer to an interpretation involving decisions about examinees as a decision-

based interpretation. These decision-based interpretations involve assumptions supporting the decision procedure's suitability as a policy, and policies are typically justified by claims about their consequences. Decision procedures that generally lead to desirable outcomes are considered sound, and decision procedures that often lead to undesirable outcomes are considered unsound. The focus is on the value judgments (what is desirable outcome?) and empirical assumptions embedded in policy assumptions.

A decision-based interpretation typically involves a descriptive component as part of the interpretation. Test scores are first given a descriptive interpretation, and then the decision is based on this descriptive interpretation. It is conceivable that the descriptive interpretation assigned to the test scores is well supported, but that the use of the test scores is inappropriate. So, it should not be very surprising if a validation effort for a basic descriptive interpretation led to the conclusion that a core set of proposed conclusions about a particular kind of achievement was valid, while a validation study examining the more ambitious decision-based interpretation suggested that the decision procedure as a whole was not valid.

The role of validation in evaluating the credibility of an interpretation per se (i.e., a descriptive interpretation) and its role in evaluating the legitimacy of a particular use (decision-based interpretation) are both recognized in the Standards for Educational and Psychological Testing:

> Validation logically begins with an explicit statement of the proposed interpretation of test scores, along with a rationale for the relevance of the interpretation to the proposed use. (AERA, APA, NCME, 1999, p. 9)

The Standards go on to say that validation involves the development of a validity argument, "to support the intended interpretation of test scores and their relevance to the proposed use." (AERA, APA, NCME, 1999, p. 9)

In cases where the decisions have serious consequences (i.e., in high-stakes contexts) the Standards clearly come down in favor of evaluating the full, decision-based interpretations, and not just the descriptive interpretations on which the decision is based:

> The test developer is responsible for furnishing relevant evidence and a rationale in support of the intended test use. The test user is ultimately responsible for evaluating the evidence in the particular setting in which the test is to be used. (AERA, APA, NCME, 1999, p. 11)

Ultimately, the intended use is to be supported by evidence.

Both Messick (1989) and the Standards (AERA, APA, NCME, 1999) include the evaluation of the consequences of test uses under validity. Those who propose to use a test score in a particular way (e.g., to make a particular kind of decision) are expected to justify this use, and these uses are generally justified by showing that the positive consequences of the proposed use outweigh the anticipated negative consequences.

Semantic Assumptions and Policy Assumptions

For purposes of this discussion, I will categorize inferences and their supporting assumptions into two broad categories, semantic and policy. Semantic inferences draw conclusions about descriptive or explanatory variables based on test scores. The semantic inferences, including for example, evaluation, generalization, extrapolation, and explanation, are those that lead from scores to conclusions or from one conclusion

to another. They make claims about what the test scores mean. I will refer to the assumptions supporting these semantic inferences as semantic assumptions.

Policy inferences lead from conclusions to decisions, and therefore involve the adoption of decision rules. The justification of such policies is generally based on claims that the decision rule will achieve certain desirable outcomes, and at least implicitly, on claims that the adoption of the decision rules will not do any serious harm. For example, calls for High-school graduation tests generally include the claim that the implementation of the testing program will lead to higher academic performance among high school students and will provide assurance that graduates have met certain standards of achievement.

The policy assumptions supporting these predictions make claims about consequences. In particular, they claim that certain positive consequences will occur because of the adoption of decision rules employing the test, that various negative consequences will not occur or at least will not be too serious, and as the bottom line, they claim that the positive consequences will outweigh the negative consequences. The evaluation of the overall consequences of a decision rule involves an application of the basic principles of decision theory (see Cronbach and Gleser, 1965), although the analysis is usually done qualitatively, because it is not feasible to explicitly define the required utility functions.

The consequences associated with high-stakes testing programs will depend on a number of characteristics of the program, including the level of the stakes (higher stakes generally make for stronger consequences, both positive and negative), the definition of the content standards on which the test is based, the characteristics of the test (e.g., multiple choice or performance), the specification of a cut score or scores,

and how the test scores are used to make decisions (e.g., alone or with other information). The policy assumptions needed to support a specific decision procedure will depend on the details of the procedure, just as the semantic assumptions needed to support a particular descriptive interpretation depend on the details of that interpretation.

An evaluation of the policy assumptions necessarily involves an evaluation of consequences. Policies are not true or untrue, accurate or inaccurate. They are effective or ineffective, successful or unsuccessful. A policy that achieves its intended goals (positive consequences) at modest cost, and with few undesirable side effects (negative consequences) is likely to be considered a success. A policy that does not achieve its goals (lack of positive consequences), and/or that involves relatively great cost or produces significant undesirable side effects (negative consequences) is likely to be considered a failure.

High School Graduation Tests

I will take as my primary example of a high-stakes test, the high school graduation tests that seem to be increasingly popular in various states. In this section, I will sketch what I take to be a reasonable interpretive argument for such testing programs, representing the assumptions that seem to be implicit in the proposed uses of the test scores. Once the interpretive argument has been specified, its inferences and assumptions can be evaluated and therefore the interpretation as a whole can be evaluated.

One of the major goals of this argument-based approach to validation, described earlier, is to make the assumptions inherent in the proposed interpretation and use of

the test scores explicit, and therefore subject to scrutiny. Some of these assumptions (e.g., assumptions about the statistical properties of test scores, about their generalizability, about their relationships to other variables) are needed to support descriptive conclusions about certain variables (e.g., that a student has achieved a certain level of skill on a certain kind of problem). Other assumptions (e.g., that certain outcomes are to be valued as the goals of assessment, and that certain costs and potential risks can be tolerated) are implicit in policy decisions.

I do not propose to evaluate whether high school graduation tests are good or bad, or valid or invalid. Mainly, I will examine the kinds of inferences and assumptions implicit in the proposed uses of high school graduation tests, and therefore, the kinds of questions that merit attention in validating the proposed uses of the test scores. The validity of a particular testing program will depend on the defensibility of the inferences and assumptions entailed by the specific interpretation/use of the test scores in that program, and this is likely to vary from one program to another. Like Haertel (1999), I am more interested in identifying the right questions than in providing general answers.

Interpretive Arguments for High School Graduation Tests

The States have a clear interest in assuring that high schools are functioning well and that the students in the state are getting a high-quality education. High school graduation tests are seen by many as an effective way to ensure that these goals are being met.

The reasoning implicit in calls for high school graduation tests seems to go something like this. It is assumed that high levels of achievement on demanding content is an important goal for high schools. It is assumed further that a core set of

desired outcomes of a high school education can be identified. Finally, it is expected that by requiring students to pass a HSGT based on demanding content, student achievement will improve.

The core set of expected outcomes is described in the state content standards, or State Standards. For example, the Wisconsin State Standards include a total of 18 Standards in English language arts, 60 Standards in science, 29 in Mathematics, and 78 in Social Studies (Fortier, Cook, and Burke, 2000).

It is claimed that the tests will focus the attention of schools, teachers, and students on the demanding content and thereby lead to higher test scores. It is also assumed that the higher test scores will reflect higher levels of achievement in the content covered by the State Standards, and that higher levels of achievement on the State Standards reflect higher overall achievement in high school. Finally, it is assumed that the benefits of the testing program will exceed the cost, where the costs include expenditures of time and money, as well as any negative side effects associated with the HSGT. Assessments based on ambitious State Standards are expected to focus attention on the attainment of these standards. High-stakes assessments are expected to focus attention more effectively.

The inferences from scores to conclusions about levels of achievement on the State Standards and to achievement in high school are semantic inferences, and their supporting assumptions are semantic assumptions. Data of various kinds can be employed to support or challenge any of these assumptions, and thereby to support or challenge the semantic inferences. Expectations about the consequences associated with HSGTs are policy assumptions, which justify the use of the tests to make graduation decisions.

It is not unusual for some of the State Standards to be excluded from the test specifications, because it is not feasible to measure them on large-scale, paper-and-pencil tests (Olson, 2001b). For example, in the Wisconsin program currently under development, 34, or 56.67%, of the 60 science standards are to be measured by the high-school graduation test (Fortier, Cook, and Burke, 2000). Standards involving specific knowledge and skills are likely to be included; standards involving more ambitious, extended performances are likely to be excluded. I will refer to the subset of the State Standards actually included in the HSGTs as the Test Standards.

We can represent the descriptive interpretation of the HSGT by a sequence of three inferences leading from scores to conclusions about achievement in high school:

I1: *From the test scores to conclusions about achievement on the Test Standards, the subset of the State Standards included in the test specifications.*

I2: *From achievement on the Test Standards (those included in the test specifications) to the full set of State Standards.*

I3: *From level of achievement on the State Standards to conclusions about overall level of achievement in high school.*

These semantic inferences are not usually stated explicitly, but they seem to be implicit in the arguments made for HSGTs. Since my purpose is to evaluate the validity of these arguments, I have tried to state them clearly and neutrally. The first two inferences are fairly straightforward. Inference 1 involves evaluation and generalization and Inference 2 involves extrapolation. Both are routinely employed in interpreting scores on achievement tests. The third inference, a second extrapolation, is harder to deal with, because any definition of the goals of high-school education is necessarily complex and highly value laden.

This is an ambitious interpretive argument leading from test scores to conclusions about overall achievement in high school. Nevertheless, even if this argument is accepted without reservation, as far as it goes, it does not justify the use of the test scores to make graduation decisions.

The justification of the decisions based on the test scores requires an additional inference/decision, a policy inference:

I4: *Students with test scores above some specified passing score are awarded a high school diploma, and students with scores below the passing score are not awarded a diploma.*

This fourth inference adds an explicit decision procedure with immediate and potentially serious consequences. It is this fourth inference that injects the "high stakes".

In the remainder of this section, I will specify the interpretive argument in more detail by outlining sets of assumptions that could be used to support each of these four inferences. I will also suggest the kinds of evidence needed to support these assumptions and some of the questions raised by each. An outline of my generic interpretive argument for high-school graduation tests, including inferences and assumptions, can be found in Table 1.

Inference 1: from scores on the test to achievement on the Test Standards (those included in the Test Specifications).

The first inference is from a sample of performances to conclusions about expected performance in the domain from which the sample is drawn. The assumptions supporting this inference and the kinds of evidence needed to support these assumptions are standard parts of most validity arguments for achievement tests.

A1.1. Each task on the test provides a reasonable measure of one or more of the State Standards.

The evaluation of this assumption will involve mainly judgement by content specialists who are expected to provide assurance that the tasks are clear and well-defined, that the scoring is appropriate and accurate, and that the tasks tap the content of the corresponding State Standard.

A1.2. The test includes measures of a representative sample of standards from the Test Standards.

The content of the test should be representative of the Test Standards as specified in the test specifications. Most of the evidence for this assumption is also likely to be based on the judgements of content specialists.

For security reasons, high-stakes tests generally involve the use of multiple forms of the test, with new forms introduced on a regular basis. (In Wisconsin, the HSGT is to be administered twice each year, with a new form for each administration.) A particular form does not need to be fully representative of the Test Standards, but the sampling of content over the multiple forms should cover all of the Test Standards. If a particular standard is seldom or never assessed, it should not be considered part of the Test Standards.

A1.3. The test scores provide dependable estimates of the expected value over the content domain defined by the Test Standards.

The dependability of generalizations from observed scores to the expected score over the universe of performances associated with the Test Standards can be examined

using a variety of reliability coefficients, generalizability coefficients, etc., based on the performance of samples of examinees (Feldt & Brennan, 1989).

Inference 2: from achievement on the Test Standards (those included in the test specifications) to the full set of Standards.

As noted above, the Test Standards are generally a subset of the State Standards. The extension of the conclusions being drawn from the Test Standards to the larger set of State Standards involves an extrapolation rather than a simple generalization, because the Test Standards are generally not a random or representative sample from the State Standards.

If the Test Standards were identical to the State Standards, this inference would not be necessary. If the Test Standards include most of the State Standards, this inference could be highly plausible a priori, and therefore, not require much support. Typically, neither of these assumptions hold.

One way to eliminate the need for this inference would be simply to take the standards actually assessed by the high school graduation test as "The Standards". This is a legitimate approach to developing an interpretive argument for the HSGT.

> When we talk about standards-based reform in Chicago, and it's actually true everywhere, don't show me the standards documents. Show me what you test," says Anthony S. Bryk, a professor of education and sociology at the University of Chicago, "because the load-bearing wall in all of this is not the standards documents, it's the assessments." (Olson, 2001a, p. 15)

This approach limits the claims that can be based on test scores to conclusions about the Test Standards (those included in the test plan) and not to the original State Standards.

In practice, it is not uncommon that, for rhetorical purposes, the scores are interpreted in terms of the full set of State Standards, but for validation purposes, the scores are interpreted in terms of the Test Standards. The inference from achievement on the test Standards to that on the State Standards is left implicit and unexamined.

Assuming that we actually want to draw inferences to the full set of State Standards, we need some basis for extrapolating expected performance on the Test Standards to expected performance on the State Standards.

A2.1. Performance on the Test Standards is positively related to performance on the full set of State Standards.

This assumption could be supported in several ways. First, it could be assumed that the Test Standards are a representative sample of the State Standards. As noted above, this assumption is generally not plausible. In many cases, the Test Standards are a substantially restricted subset of the State Standards.

Second, it could be claimed that scores on the test are closely related to performance on the State Standards, in spite of not being a random or representative sample. Since the Test Standards are a subset of the State Standards, all else being equal, performance on the Test Standards should be positively related to performance on the State Standards, with the strength of the relationship depending to some extent on the overlap between the Test Standards and the State Standards.

This approach to justifying assumption, A2.1, would be much less plausible, a priori, for a high stakes test (e.g., HSGT) than it would be for a low-stakes test (e.g., NAEP, some state testing programs). One of the basic arguments for high-stakes tests, like the HSGT, is that they focus attention on demanding content (i.e., on the test

content). To the extent that this focusing is effective, we may see improvements in the test content with little or no improvement in other areas. In fact, if the focusing effect is strong enough, it is possible that less attention would be given to other areas of the curriculum (including those State Standards that are not included in the Test Standards), as more attention is given to the Test Standards. In the extreme case, achievement on the Test Standards could improve, while achievement on the full set of State Standards declined, because the State Standards not included in the Test Standards are being ignored. This extreme scenario seems unlikely, but not impossible.

Assuming that test scores based on a modest set of Test Standards are used to draw conclusions about a more ambitious set of State Standards, it would seem imprudent to take the relationship between achievement on the Test Standards and achievement on the full set of State Standards for granted.

This relationship could be evaluated by having some sample of students assessed on the full set of State Standards, and separately, on the HSGT, and then comparing scores on these two measures, thus yielding criterion-related validity evidence. It could also be evaluated by conducting systematic studies, within some sample of schools, of the extent to which the increased emphasis on the Test Standards engendered by the HSGT leads to decreased emphasis on those State Standards not included in the HSGT.

Inference 3: from achievement on the State Standards to conclusions about overall achievement in high school.

As was true for the last inference, Inference 3 can be justified in several different

ways, with the differences in the assumptions made leading to differences in the shape of the interpretive argument and therefore in the interpretation.

As was the case for Inference 2, the easiest way to justify this inference is by fiat. The need for Inference 3 and for the evidence to support it disappears, if the State simply declares that, for the purposes of the high school graduation test, the State Standards define the scope of desired outcomes of high school education. In terms of graduation decisions, those things not included in the State Standards (e.g., sports, art, music, PE, foreign languages) would be considered largely irrelevant. This approach might not be political palatable, but I think that it would be legal. One version of Catch-22 said that, "They have a right to do anything that you can't stop them from doing." (Heller,)

Alternately, the State Standards could be is assumed to be a representative sample from the full range of goals of a high school education. The evaluation of such a claim immediately leads to the question of how to define the goals of high school (Eisner, 2001). Since the State Standards focus on cognitive skills, and, in fact, a limited range of such skills, they would provide a fairly narrow conception of the goals of HS. They generally do not include physical skill and health-related physical fitness, personal qualities (e.g., leadership, perseverance, and integrity), skill and creative expression in the arts, facility with critical thinking and problem solving on real-world tasks (e.g., scientific experiments, practical problems), and useful practical skills like driving.

A somewhat more plausible assumption would recognize that the State Standards contain only a subset of the goals of secondary education, but would claim that it is an important subset and that achievement of the other goals will be positively

correlated with achievement on the State Standards.

A3.1. Achievement on the State Standards provides a good indication of overall achievement in high school.

The a priori plausibility of this assumption depends to a large extent on the breadth of coverage of the State Standards and on how one defines the expected outcomes of a high school education. If we define expected goals of high school broadly to include things like musical accomplishment, artistic skill and expression, physical fitness, group problem-solving skills, and character, this assumption seems less plausible, a priori, than it would be if the goals of a high-school education were limited to mastery of basic skills in the core academic subjects. For example, Eisner (2001) suggests that developing the ability to raise telling questions should be a major goal of education:

> Are students encouraged to wonder and to raise questions about what they have studied? Perhaps we should be less concerned with whether they can answer our questions than with whether they can ask their own. (Eisner, 2001, p370)

In standardized testing, students are expected to answer question rather than ask them.

The relationship between achievement on the State Standards and overall achievement in high school could be evaluated empirically, by assessing a sample of students on some measure of overall level of achievement in high school (in academics more generally, in sports, in community activities, etc.) and relating these results to scores on the HSGT. As was the case with assumption A2.1, the plausibility of assumption A3.1 could also be evaluated by examining the extent to which the emphasis on the State Standards and the HSGT leads to diminished emphasis on valued content and activities not covered by the State Standards. The extent to which

these tradeoffs are acceptable is a matter of policy, and will ultimately be decided by the policy makers and by groups of stakeholders who can influence the policy makers.

Assumption A3.1 would also clearly be more plausible in a low-stakes context than in a high-stakes context. To the extent that the high-stakes are successful in focusing the attention of teachers, administrators, parents, and students on the State Standards, and more specifically on the HSGT, other aspects of high-school education may get less attention.

Our expectations of a high-school education could have a profound impact on individual students and on our society, and should probably not be adopted implicitly and casually. Yet, this issue seems to have received remarkably little attention. In our enthusiasm for high standards on demanding content, we have generated some very impressive sets of State Standards. However, I have not seen much critical discussion of the appropriateness of different kinds of content standards for a high-school graduation test. Presumably, these standards should reflect a level of achievement that is necessary for some purpose. If the State is going to deprive a student of a diploma after they have been required to spend twelve years in school, there should presumably be some reason for doing so (other than an abstract desire for "world-class" standards). Presumably, the State Standards represent content and performance standards that are thought to be required for success in some future endeavor? If so, what is the endeavor? Is it college, technical school, work, the military life in general? In a discussion of Texas HSGT, Ward (2000) argues that

> The more expansive interpretation sees successful performance on a standards-based graduation test as a way of demonstrating to the public, the business community, postsecondary educational institutions, etc., that, "students are leaving high schools prepared with skills to function in today's world." (Ward, 2000, p. 421)

It seems to me that the content and performance standards appropriate for students entering the university are potentially different from those that are appropriate for students going to work directly out of high school.

The State Standards in Wisconsin seem to reflect a traditional college-preparatory curriculum, with mathematics, science, social studies, and language arts. Is this an appropriate way to operationalize our collective vision of a good high-school education? Should we adopt a different model, or a range of different models to accommodate different patterns of interests and talents among high-school students?

***Inference 4:* from conclusions about overall achievement in high school to a decision about whether to award a HS diploma.**

The advocates of HSGTS see these tests as a way to achieve certain goals, in particular, to raise academic standards by focusing attention on demanding content, (presumably, while not having a negative impact on any of the other desirable outcomes of a high-school education). According to Philips (2000), the Texas Education Agency has decided that the purpose of the Texas high school graduation test is, "to ensure a minimal level of competence in both mathematics and reading." (p. 371), and the judicial opinion which sided with the TEA's point of view, decided that it is appropriate to use the test, "to hold students, teachers, and schools accountable for learning and teaching, to ensure that all students have the opportunity to learn minimal skills and knowledge, and to make the Texas high school diploma uniformly meaningful" (Philips, 2000, p. 371).

The use of the HSGT to promote achievement in certain content areas is clearly

an educational policy, much like the adoption of a new curriculum. The focus is not on measurement per se, but on achieving certain goals. The test is being used to implement a policy. This policy will have consequences, some of which are likely to be positive and some negative. The test-as-policy should be evaluated in terms of its perceived effectiveness.

The usual argument in favor of this policy is based on two assumptions, that the adoption of the HSGT will lead to improvements in the areas tested, and more generally, in the areas defined by the State Standards, and that this will be accomplished without serious losses in other areas.

A4.1: High School Graduation Tests (HSGT) will lead to higher levels of achievement on the State Standards.

By focusing attention on the State Standards, the HSGT requirement is to promote higher levels of achievement on the content covered by the State Standards.

One source of evidence supporting this assumption could be derived from longitudinal studies of performance on the HSGT, along with evidence (collected in evaluating inferences 1 to 3, and their supporting assumptions) relating the changes in scores on the HSGT to conclusions about changes in achievement on the State Standards and on the broader conceptions of achievement in high school.

Evidence relevant to this assumption could also be derived for studies of changes in the schools after the HSGT is introduced. Are more students taking more demanding courses (e.g., algebra, physics) and is the content of all courses becoming more rigorous? Is the number of students taking advanced placement courses increasing?

The central question here is the extent to which the introduction of the HSGT changes patterns of effort and achievement on the State Standards for the better. Therefore, the research needed to examine this assumption will have to involve pre-post designs or longitudinal designs, and should begin to collect baseline data before the introduction of the HSGT. The questions to be addressed are essentially those traditionally asked in the context of program evaluation. This seems appropriate, because the HSGTs are being used as educational interventions and not simply as measurement instruments. The descriptive part of the interpretation (i.e., I1, I2, and I3) can be evaluated using traditional approaches to validation, but the decision-based part (i.e., I4) involves the effectiveness of certain policies, and the evaluation of these decision procedures requires experimental, or more likely, quasi-experimental methods.

A4.2: The adoption of the HSGT will not have a major negative impact on achievement in other areas.

This assumption basically claims that the implementation of the HSGT will not cost much beyond the resources needed to develop and administer the test. Evidence for this assumption can be generated by examining the unintended effects of the HSGTs. Does the high-school dropout rate increase after the introduction of the test? Does the number of students taking art, music, physical education decrease? Does participation in extracurricular activities (sports, clubs) go down? Are students spending a lot of time and effort on test-preparation activities?

The rhetoric of the new accountability movement is quite similar to that of the minimal competency movement of the early 80's. The difference is that the tests have been made more demanding and the performance standards are higher. In discussing

the standard setting for the Texas graduation test, Philips argues that:

> The earlier graduation test focused on basic skills; the newer graduation test covered the same curricular areas but placed more emphasis on higher order thinking and problem-solving skills. Thus, by design, the 1990 graduation test was more difficult than its predecessor. (Phillips, 2000, p. 359)

The main criticism of the earlier basic-skills tests was that the tests led to an emphasis on basic skills at the expense of higher order thinking and problem solving. The question that needs to be addressed for the new testing programs is what if anything they are pushing off the stage.

Evaluating Policy Assumptions: Costs/benefit Analyses

In evaluating the costs and benefits of the testing program for students, it will probably be necessary to recognize that the outcomes are likely to be quite different for different groups of students. For example, we might consider three broad categories of students, high achieving students who pass the test on their first try, low achieving students who fail the test several times (and perhaps never pass), and marginal students who get scores close to the passing score. I would expect both the benefits and costs to be very different for these three groups. In particular, the potential benefits and unintended effects are likely to be most pronounced for the marginal group, especially for the surprises - students who are doing fairly well in school but fail the test one or several times, and students who have not been doing well in school, but pass the test on their first or second try. The high school graduation test could be a life altering event for some such students.

Heubert and Hauser (1999) report on research by Griffin and Heiden (1996) suggesting that the students most likely to drop out after failing a high school graduation

test are those who were doing fairly well academically. Griffin and Heiden (1996) also report that students with poor academic records did not seem to be as bothered by failing the test. On the other hand, marginal students who pass the HSGT may have an added incentive to complete high school. If they complete their courses with passing grades, they can have confidence that they will get a diploma. Which of these scenarios is likely to occur most often?

The expected benefits of high-school graduation tests are many, and include the following:

1. Overall student achievement on the State Standards may improve.
2. Overall student attainment of the larger set of legitimate goals of secondary education may improve.
3. The adoption of the HSGTS is supposed to provide assurance that all high school graduates have attained a fairly high level of achievement on the State Standards.
4. Students may be encouraged to take more demanding courses.
5. The content of various courses may become more demanding.

On the other hand, the critics (Eisner, 2001; Thompson, 2001) have pointed to a number of potential negative effects of HSGTs in addition to the direct, anticipated costs of the testing program:

1. Dropout rates may increase
2. Participation in sports and extracurricular activities may decrease.
3. Schools may have to drop electives (e.g., languages, AP courses) to provide resources to remedial courses.
4. Students may choose to take fewer electives.

5. Costs in terms of time and money involved in the development, administration and scoring of the tests, as well as the costs of potential litigation are likely to be quite substantial.

To the extent that these side effects occur, their impact is likely to be different in different groups.

In evaluating the appropriateness of test use, it is the consequences of the decisions that count. The "bottom line" requirement is that expected benefits should outweigh the expected costs. Cronbach and Gleser (1965) laid out the rules of accounting for expected costs in a formal way for selection and placement decisions, using utility theory. Unfortunately, it is not possible to carry out this kind of precise analysis of the outcomes of a HSGT.

It is possible, however, to conduct some analyses of the outcomes, positive and negative, associated with the testing program. It is possible to estimate the extent to which the intended outcomes of the HSGT actually occur in the high schools (e.g., greater emphasis on demanding content, specified in the State Standards) and in individual students (e.g., higher levels of achievement on the State Standards). It is also possible to estimate the extent to which specific unintended outcomes (e.g., increased dropout rates, decreases in elective course offerings and enrollment, declines in participation rates in extracurricular activities) occur after the introduction of the HSGT. The changes in all of these outcomes can be examined as a function of various demographic variables and of the academic standing of the students (e.g., high, marginal, low).

The overall value of the HSGTs can then be determined by weighing the impact of the testing program on these outcomes. This would presumably be done by those

responsible for setting educational policy, with input from stakeholder groups.

CONCLUSIONS

Cronbach (1988) has pointed out that for validity arguments to be convincing to diverse audiences, the assumptions in these arguments must be credible to those audiences. This concern is especially salient for the policy assumptions implicit in the arguments for high-stakes testing programs, because the credibility of these policy assumptions may be highly variable across different groups of stakeholders.

In this paper, I have tried to draw a distinction between descriptive interpretations and decision-based interpretations, and a related distinction between semantic assumptions and policy assumptions. Descriptive interpretations draw conclusions about a student based on the student's performance on the test, but do not involve any decisions. The decision-based interpretations include decision procedures and therefore rely on policy assumptions to a much greater extent than the descriptive interpretation.

To adopt the practice of making a certain kind of decision for a population using a test is to adopt a policy. The decisions made with high-stakes assessments are important and therefore the policy assumptions built into these assessment systems are important. Yet, it is easy for these policy assumptions to be obscured by the surrounding technical framework so that the nature and implications of the policy assumptions being made is not clearly understood by anyone, perhaps least of all by the public and public officials.

I think that much of the current practice in the validation of high-stakes testing programs including HSGTs is seriously flawed, because only a part of the interpretive

argument is evaluated. The interpretations that are assigned to the results of these tests are very ambitious, extending from the scores to conclusions about achievement on the Test Standards, to conclusions about achievement on the State Standards, and then to conclusions about achievement in high school. Finally, the results are used to make decisions about the awarding of high-school diplomas. The adoption of these decision procedures is claimed to have a variety of positive systemic effects, particularly in terms of focusing the attention of high-school administrators, teachers, and students on demanding content.

Yet, the validity arguments developed to support these ambitious claims typically attend only to the initial steps in the interpretive argument. The validity evidence that is provided tends to emphasize the inferences from the test scores to achievement on the Test Standards. The additional semantic inferences to achievement on the State Standards and to conclusions about overall achievement in high school are simply taken for granted. I have not seen any indication that any state is seriously investigating the policy assumptions inherent in Inference 4, the decision. These policy assumptions are also taken for granted.

In logic, this kind of faulty argument, or fallacy, has a name. It is referred to as, "begging the question," because the question to be decided, or a large part of the question to be decided, is simply taken for granted (Walton, 1989). In the case of the HSGTs, the inference from the test scores to achievement on the Test Standards is supported by evidence (generally content-related evidence), and we are asked to take the rest of the interpretive argument as given.

I am not arguing that the proposed uses of high school graduation tests are invalid. I don't know if the assumptions being made are or are not valid. The trouble is

that neither does anyone else, although everyone has an opinion. I do think that it would be prudent to investigate the assumptions built into the interpretations the HSGTs, before we subject a generation of high-school students to a potentially powerful treatment.

> Test validation includes collecting evidence on the intended and unintended consequences of test use. Determining whether the use of a test for making graduation decisions produces better overall educational outcomes requires that the various intended benefits of test use be weighed against unintended negative consequences for individual students and different kinds of students (American Educational Research Association et al., 1985: Standard 6.5; Joint Committee on Testing Practices, 1988; Messick, 1989). (Heubert & Hauser , 1999, p. 172)

The evaluation of the policy assumptions inherent in high-stakes testing programs raises some difficult issues, but they are not any more difficult than the issues routinely faced by program evaluators. And, in fact, if the primary purpose of a testing program is to promote certain outcomes, rather than simply to estimate certain variables, the testing program is functioning as an educational intervention, and therefore merits evaluation of the kind routinely mandated for any new educational program.

For stakeholders to make informed decisions about the effectiveness of HSGTs, it is necessary that they have information about how well these tests perform in promoting various goals and at what cost. To get this information, it will be necessary to conduct the kind of research typically found in program evaluations. We need to collect baseline data before the HGSTs are implemented and then examine trends, positive and negative, after the tests are introduced. Assuming that there are both positive and negative consequences, the stakeholders and policymakers face the task of evaluating the overall consequences.

Table 1.

Outline of Interpretive Argument for HSGTs

Inference 1 from the test scores to conclusions about achievement on the Test Standards (those included in the test specifications).

A1.1. Each task in the test provides reasonable measures of one or more of the State Standards.

A1.2. The Test includes measures of a representative sample from the Test Standards.

A1.3. The test scores provide dependable estimates of the expected value over the content domain defined by the Test Standards.

Inference 2 from achievement on the Test Standards (those included in the test specifications) to achievement on the full set of State Standards.

A2.1. Performance on the Test Standards is positively related to performance on the full set of State Standards.

Inference 3: from achievement on the State Standards to conclusions about overall achievement in high school.

A3.1. Achievement on the State Standards provides a good indication of overall achievement in high school.

Inference 4: from conclusions about overall achievement in high school to a decision about whether to award a HS diploma.

A4.1: The Adoption of a High School Graduation Test Requirement (HSGT) will lead to higher levels of achievement on the State Standards.

A4.2: The adoption of the HGT will not have a major negative impact on achievement in other areas.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Bishop, John. H. (1997). The effect of national standards and curriculum-based exams on achievement. American Economic Review, 87(2), 260-264.

Brennan, R. L. (1992). Elements of generalizability theory, Revised edition. Iowa City, IA: American College Testing.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), Educational measurement, 2nd ed. (pp. 443-507). Washington, D.C.: American Council on Education.

Cronbach, L. J. (1980a). Validity on parole: How can we go straight? New directions for testing and measurement: Measuring achievement over a decade. Proceedings of the 1979 ETS Invitational Conference (pp. 99-108). San Francisco: Jossey-Bass.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), Test validity (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J., & Gleser, G. C. (1965). Psychological tests and personnel decisions. Urbana, IL: University of Illinois Press.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley.

Crooks, T., Kane, M. & Cohen, A. (1996). Threats to validity. Assessment in Education, 3, 265-285.

Eisner, E. (2001). What does it mean to say that a school is doing well?. Phi Delta Kappan, January, 367-372.

Fortier, J., Cook, H., and Burke, M. (2000) Wisconsin High School Graduation Test: Educator's guide. Wisconsin Department of Public Instruction, Madison, WI.

Green, D.R. (1998). Consequential aspects of the validity of achievement tests: A publisher's point of view. Educational Measurement: Issues and Practice, 17, 2, 16-19, 34.

Griffin, B.W. & Heidorn, M.H. (1996). An examination of the relationship between minimum competency test performance and dropping out of high school. Educational Evaluation and Policy Analysis, 18(3), 243-252.

Haertel. E. (1999). Validity arguments for high-stakes testing: In search of the evidence. Educational Measurement: Issues and practice, 18, 4, 5-9.

Haney, W. (1991). We must take care: Fitting assessments to functions. In V. Perrone (Ed.), Expanding student assessment. Arlington, VA: ASCD.

Heller, J. (1955). Catch 22, S&S Classical Edition. Simon and Shuster, New York.

Heubert, J.P. & Hauser, M. H., (1999) High stakes: Testing for tracking, promotion, and graduation. Nation Academy Press, D.C.

House, E. R. (1980). Evaluating with validity. Beverly Hills, CA: Sage Publications.

Kane, M. (1992). An argument-based approach to validation. Psychological Bulletin, 112, 527-535.

Kane, M. T., Crooks T.J., & Cohen, A.S. (1999). Validating measures of performance. Educational Measurement: Issues and Practice, 18, 2, 5-17.

Linn R. L. (1997). Evaluating the validity of assessments: The consequences of use. Educational Measurement: Issues and Practice, 16, 2, 14 - 16.

Linn, R.L. (1998). Partitioning responsibilty for the evaluation of the consequences of assessment programs. Educational Measurement: Issues and Practice, 17, 2, 28-30.

Mehrens, W. A. (1997). The consequences of consequential validity. Educational Measurement: Issues and Practice, 16,2, 16 - 18.

Mehrens, W.A. (2000). Defending a state graduation test: *GI Forum V. Texas Education Agency.* Measurement perspectives from an external evaluation. Applied Measurement in Education, 13(4), 387-401.

Messick, S. (1988). The once and future issues of validity. Assessing the meaning and consequences of measurement. In H. Wainer and H. Braun (Eds.), Test validity (pp. 33-45). Hillsdale, NJ: Lawrence Erlbaum.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational Measurement, 3rd ed. (pp. 13-103.) New York: American Council on Education and Macmillan.

Moss, P. (1993). Shifting conceptions of validity in educational measurement: Implications for performance assessment. Review of Educational Research, 62, 229-258.

Moss, P.A. (1998). The role of consequences in validity theory. Educational Measurement: Issues and Practice, 17, 2, 6-12.

Newmann, F.M. (1992). The assessment of discourse in social studies. In H. Berlak, F.

Newmann, E. Adams, D. Archbald, T. Burgess, J. Raven, and T. Romberg, Toward a New Science of Educational Testing and Assessment (pp. 53-69). Albany, State university of New York Press.

Newmann, F.M. & Archbald, D.M. (1992). The nature of authentic academic achievement. In H. Berlak, F. Newmann, E. Adams, D. Archbald, T. Burgess, J. Raven, and T. Romberg, Toward a New Science of Educational Testing and Assessment (pp. 71-83). Albany, State university of New York Press.

Ohanian, S (2001). News from the test resistance trail. Phi Delta Kappan, January, 363-366.

Olson, L. (2001a). Finding the right mix. Education Week, Vol. XX, No. 17, Jan 11, 2001, 12-20.

Olson, L. (2001b). Overboard on testing. Education Week, Vol. XX, No. 17, Jan 11, 2001, 23-30.

Phillips, S.E. (1991). Diploma sanction tests revisited: New problems from old solutions. Journal of Law and Education, 20(2), 175-199.

Shepard, L.A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), Review of Research in Education, Vol. 19 (pp. 405-450). Washington, DC: American Educational Research Association.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. Educational Measurement: Issues and Practice, 16, 2, 5-8,13,24.

Taleporos, E. (1998). Consequential validity: A practitioner's perspective. Educational Measurement: Issues and Practice, 17, 2, 20-23, 34.

Thompson, S. (2001). The authentic standards movement and its evil twin. Phi Delta Kappan, January, 358-362.

Walton, D. (1989). Informal logic: A handbook for critical argumentation. Cambridge: Cambridge University Press.

Ward, C. (2000). GI Forum v. Texas Education Agency: Implications for state assessment programs. Applied Measurement in Education, 13, 419-426.

Yen, W.M. (1998). Investigating the consequential aspects of validity: Who is responsible and what should they do? Educational Measurement: Issues and Practice, 17, 2, 5.

TM032854

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: THE ROLE OF POLICY ASSUMPTIONS IN VALIDATING HIGH-STAKES TESTING PROGRAMS

Author(s): MICHAEL KANE

Corporate Source: UW MADISON

Publication Date: APRIL, 2001

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1 ☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A ☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B ☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: Michael T. Kane	Printed Name/Position/Title: MICHAEL T. KANE / PROF
Organization/Address: UNIV. of WISCONSIN MADISON, MADISON, WI 53706	Telephone: (608) 265-2891 FAX: 608-262-1656
	E-Mail Address: KANE@EDUCATION.WISC.EDU Date: 4/24/01

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

EFF-088 (Rev. 9/97)