DOCUMENT RESUME

ED 204 380 TM 810 370

AUTHOR Beck, Michael D.
TITLE Uses and Misuses of Standardized Test Scores on a Local Level--A Test Developer's Perspective.
PUB DATE Apr 81
NOTE 10p.: Paper presented at the Annual Meeting of the American Educational Research Association (65th, Los Angeles, CA, April 13-17, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Elementary Secondary Education: *Local Norms: Norm Referenced Tests: *Scores: *Standardized Tests: *Testing Problems: *Test Interpretation
IDENTIFIERS *Test Use

ABSTRACT

Standardized test scores (STS) should be used on a local level: (1) as one component of evaluation of a student, school, or district: (2) to draw as much interpretive meaning from a norm-referenced test (NRT) as their structure will support: (3) as a communication device with students, parents, the public, and professional staff: (4) to check status across grade levels in key subject matter areas: (5) to compare inter- and intra-grade scores across years: (6) to discover individuals whose measured achievement deviates significantly from their school ability level: and (7) to compare achievement scores with national, local, and other subgroup normative data sets. The misuse of STS is outlined, as are the several weaknesses or limitations of using "local norms" based only on the district population. Local norms often depend on unqualified local personnel for their interpretation, should be developed once and then re-used over the next few years in order to track year to year changes, and may lead to educationally damaging interpretations in low-performing districts. (RL)

USES AND MISUSES OF STANDARDIZED TEST SCORES ON A LOCAL LEVEL -- A TEST DEVELOPER'S PERSPECTIVE

Michael D. Beck
The Psychological Corporation

Standardized tests have a large number of uses and an even larger number of misuses on a local level. The following uses and misuses strike me as being especially important - or, at least, interesting to discuss. It is important to stress at the start that I do not pretend to speak for my colleagues, either within The Psychological Corporation or in the "industry" in general. These are merely my biases; others should feel free to disagree, if they choose to be wrong.

The uses and misuses I wish to address can be conveniently grouped into three major groups - testing's proper role in decision making, using tests to assess status and change, and interpreting results in terms of various frames of reference.

I. TEST'S PROPER ROLE

Use: As one component of evaluation of a student, school, or district.

Misuse: As the sole criterion for decision-making: e.g., minimum competency tests, promotion decisions, Title I evaluations, teacher evaluation.

No important decision should ever be made based on a single piece of information, no matter how "reliable and valid" the information is. Yet, critical educational decisions are made daily about children, teachers, programs, and schools based solely on single isolated sets of test scores. No doubt, part of the blame for this rests with our enamor for numbers. Yet, that is a simplistic and incomplete explanation. I believe a larger part of

---

Paper presented at the meeting of the American Educational Research Association, Los Angeles, April 1981.

the problem here is a sad reality -- we (collectively) have simply failed to use anything other than norm-referenced tests to evaluate schools or kids. How can we assail newspaper reporters, boards of education, legislators, "the public" for overinterpreting test scores when we provide no other evidence? Isn't it time we all got serious about seeking out sound alternative assessment devices and using them -- in addition to, if not instead of standardized tests? If we all honestly believe that a 40-minute reading test provides a less complete answer to our questions "Can Jill read?" or "Did the program work?" than does a trained professional, let's all say so and provide the other evidence. When we confront untrained people with a single piece of data, expecially one with decimalized numbers, is it really surprising that judgments are based solely on that one piece of data?

All of that said, I'd like also to mention that one piece of data is at least a step in the right direction. The NEA ostrich position provides little solace to a disgruntled, skeptical public. NEA's argument is perceived -- correctly in my unbiased eyes -- as being essentially: "One piece of information isn't sufficient for decision-making, especially when the information is far from perfect. Therefore, let's not provide any information and there won't be any problems." Fortunately, few people -- even among NEA's own membership -- have been able to follow the logic any better than I. However, anyone who believes educational accountability is passé just because it's the topic of only one AERA session this year has simply not been listening to people in the real world.

Use: Drawing as much interpretive meaning from a NRT as their structure will support.

Misuse: Pretending NRTs are "diagnostic" and criterion-referenced.

Regardless of what publishers' advertisements and promotion brochures claim, or what all of us would like, one test cannot be all things to all people. The test with greatest market appeal today would be a twenty-minute diagnostic/prescriptive basic skills achievement battery -- with objective mastery cutoffs and, of course, a full complement of normative data. Sooner or later, we -- all of us in this game -- are going to have to honest and tell some people we don't have such an instrument and we never will.

There is a fine, and often indistinct, line between milking a test for as much information as it can reasonably yield and over-interpretation. My personal bias is that all of us have, more often than we like to admit, crossed the line.

Use: As a communication devide -- with students, parents, "the public" and professional staff.

Misuse: Failure to report results to all concerned groups.

A recent national survey (Beck & Stetz, 1979) indicated that while almost 90% of the students in Grades 5-12 would like to find out their scores on standardized tests, fewer than 1/3 of their teachers report the scores to students. Is it surprising when we hear of students whose attitudes toward taking such tests are less-than-ideal or who give less than 100% to completion of the task? If you were told yearly that the test you were taking was important, would be used to help you, and that you should do your best, how seriously would you take this information after being lied to five or seven times? On a broader level, if results are not routinely shared with parents, the public, and the staff, shouldn't we expect results to be viewed with suspicion or distrust or as not useful?

II. STATUS & CHANGE

Use: Checking status across grade levels in key subject matter areas
Comparing inter- and intra-grade scores across years.

Misuse: Smorgasbord testing programs. Frequent changes in test series.

Probably the single most useful attribute of standardized, norm-referenced test scores is their comparability within grade across content areas and within content areas across grades. That is, such scores -- probably uniquely -- permit schools to assess (in a normative sense at least) relative status across subject matter areas and across grades.

Nevertheless, a disturbingly large percentage of school districts are unable to avail themselves of this feature. Why? For whatever reasons -- budget, poor leadership, a committee approach to decision-making, not wanting to disappoint any publisher -- sizable numbers of districts use 2, 3, or more series at different grades in any given year. Such a testing policy, by definition, eliminates one of the potentially most useful features of any standardized test. When such a policy exists, the school either is unable to look at changes across grades/content areas or, worse, "looks" at them but draws unsupportable, totally inaccurate conclusions. At the risk of slight overstatement, I doubt that there is ever an educationally sound justification for such a smorgasbord testing plan.

A related, though perhaps less pervasive, problem is that of changing a systemwide test series on a frequent basis. "Frequent" is difficult to define, though for discussion purposes, I would be hard pressed to support changes more frequent than every four or five years. Even when changes are made from an old to a newer edition of a test series, the problem is present -- despite what publishers' equating tables would indicate. For many test uses,

the consistency of the norms is of greater importance than their accuracy vis à vis some theoretical national average. Each time a test series is changed, this consistency is lost. The interpretive value of an unchanging frame of reference is often overlooked in the search for the most up-to-date test and test norms as possible. In most instances, consistency of the norms has much greater import than does currency.

Use: Assessing change -- individual group -- over time.

Misuse: Unrealistic "growth" expectancies.

One of the primary reasons schools use NRTs is to assess change: In fact, perhaps the broadest "use" of such tests today is in Title I and other compensatory programs, in which change assessment is the primary, if not sole, purpose. Nevertheless, there continue to be large numbers of districts in which growth expectancies are totally unreasonable.

Examples of this situation are not difficult to find. They include the following, each of which occurs frr more frequently than any of us -- publishers, "informed" users, evaluators, ivory-tower academicians -- would like to admit:

"All students in this program should show 5 NCE units' growth."

(This is the 1980's version of the "year's growth for a year's instruction" slogan. Hard as this is for our DOE, RMC, TAC and other alphabet friends to accept, the current slogan is only marginally more digestible than its distasteful predecessor.)

"The average PR for each of our elementary buildings will increase by 10 points this year."

"Every child (building) should score above average on this test."

"37% of our students failed to show normal growth this year."

III. FRAMES OF REFERENCE

Use: Comparing ability and achievement test results to discover individuals whose measured achievement deviates significantly from their school ability level.

Misuse: Interpreting small, non-significant ability-achievement differences as revealing problems. Comparing results on ability and achievement tests that were not normed together. Considering ability test results as indicating innate, immutable intelligence.

Interpreted with the appropriate amount of caution, analysis of significant differences between concurrently normed ability and achievement tests can be revealing and instructionally useful. The key portions of this position are "with the appropriate amount of caution," "significant differences," and "concurrently normed." If any of these are not met in a specific instance, the value of the comparisons will range from meaningless to harmful. A subtle but, I believe, meaningful distinction in ability-achievement comparisons is between interpreting the results in an "expectancy" sense versus in a "predictive" sense.

Many of my more academically oriented colleagues continue to prolong the age-old debate of whether intelligence/ability tests actually measure anything distinct from achievement tests. It's really time to put this silly topic to rest -- <u>of course</u> they measure different things. Not totally different, not uncorrelated, not two sets of unique characteristics, but clearly different things. Of course how able someone is relates slightly with his current achievement. And of course what someone has learned to date affects how able she is to learn future things. But to pretend that ability and achievement are one and the same -- or that even current state-of-the-art measures assess the "same thing" despite their labels is patently false and inattentive to facts.

Looked at purely in a statistical sense -- which I guess AERA meeting presenters should do -- assume a typical pair of achievement and ability tests. The ability test would have a reliability coefficient of about .90 -- most are better, but I work better with rounded numbers. All of us technicians visualize a pie called intelligence or ability or some such in which 90% is "clean" (whatever it is we're really measuring) and the other 10% is garbage -- "error" to you purists. Now let's add the achievement test. The typical achievement-ability test correlation is about .75. Using my psychometric snake oil, I come up with a new ability-achievement pie in which we still have the 10% garbage, a 56% slice called achievement-ability, and 34% that's unique to the ability test. No one can honestly claim that something that accounts for over a third of the pie is trivial. Not something we <u>want</u> to pay attention to for whatever reason -- time, politics, cost -- OK. Not 'worth it" in a measurement sense -- short-sighted.

Use: Comparing achievement scores with appropriate benchmarks -- national, local, other subgroup normative data sets.

Misuse: Selecting benchmarks that result in misleading conclusions about status or change.

National norms, despite what the popular press would have us believe, are <u>not</u> on their way out. Such data continue to be widely requested and almost-as-widely used. I see no signs of the demise of national norms for the traditional types of survey tests.

On the other hand, other types of normative data <u>are</u> frequently requested and, far less frequently, provided for certain NRTs. I'm thinking here of such subgroup data as regional, public vs. non-public, large-city, Title I, special education, and socioeconomic status norms. Many, if not most, test

users would like to have some type of sub-national norms for the tests they are using. Many of the norms sets currently being provided for such purposes, are remarkably devoid of technical soundness, however. Potential users of such norms sets need to inspect the representativeness of these data very carefully rather than faithfully adopting the data as if they were sound and well-developed.

A final set of popular norms for tests is "local norms," based only on the district population. Despite their surprisingly wide use, and sanction by most measurement specialists, these data have several weaknesses or limitations:

1) It is extremely rare to find local personnel who can correctly interpret such data. I have repeatedly heard such things as, "National norms tell how we compare to people nationally and local norms compare us with similar local districts." Or, "Our district average is at the 43rd percentile in national norms, but we're right at the 50th in terms of local norms." Or, "Our averages in national norms have been dropping over the past few years, but with local norms, we're holding our own." Such statements give me little reason to suspect that local norms are interpreted nearly as well as are national norms. And we all know how well national norms are interpreted.

2) In order to be most useful, local norms should be developed once and then re-used over the next few years. Otherwise, year to year changes cannot be tracked. However, I know of no large-scale development and use of local norms of this type.

3) In low-performing districts, local norms can lead to educationally damaging interpretations. The real message often conveyed in such cases is, "Mrs. Jones, your son can't read, but he can read almost as well as the other kids who can't read." This is, of course, not a problem with local norms, per se, but with their interpretations.