DOCUMENT RESUME

ED 222 576 TM 820 741

AUTHOR Haney, Walt
TITLE What Could Be Done Differently with NAEP?
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE Aug 82
NOTE 16p.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Change Strategies; Cost Effectiveness; *Educational Assessment; Evaluation Methods; Federal Programs; Needs Assessment; Planning; Policy Formation; Program Design; *Research Utilization; Testing Programs; *Test Interpretation; *Test Use; Test Validity
IDENTIFIERS *National Assessment of Educational Progress; National Institute of Education

ABSTRACT

A needs assessment of the National Assessment of Educational Progress (NAEP) is presented. It deals with cost, design and technical issues, and utility. Suggestions include cost reduction via assessment schedule cutbacks and re-use of released NAEP exercises; and a shift from federal to private funding by selling NAEP exercises with interpretative materials to schools or individuals. A major unresolved question concerns the validity of inferences which can be drawn from the aggregated results. NAEP validation procedures constitute content validation. However, content validation does not necessarily constitute validity evidence at all, if validity evidence must bear on the interpretations that are warranted on the basis of test or assessment results. This means that more work needs to be done on the construct validation of the NAEP results as they are now commonly interpreted, i.e., with results aggregated across exercises. NAEP needs to become more clear in reporting aggregate results by specifying the facets of variance over which it is, and is not, attempting to generalize. Two strategies are suggested for the utility of NAEP results: developing norms for NAEP exercises, and making NAEP exercises and data more readily available to independent investigators. (PN)

WHAT COULD BE DONE DIFFERENTLY WITH NAEP ?

Walt Haney
The Huron Institute
123 Mt. Auburn St.
Cambridge, MA 02138

August 1982

This brief paper has been written at the request of the National Institute of Education. It has been developed as part of a broader examination by the author comparing alternative sources of national evidence on student learning. Ideas and suggestions advanced in the paper do not of course necessarily reflect those of anyone other than the author.

HANEY

WHAT COULD BE DONE DIFFERENTLY WITH NAEP?

In this brief memo, I set out a range of ideas for what could be done differently with and for NAEP. Ideas are organized roughly into three categories, pertaining to cost, design and technical issues, and utility. I do not discuss administrative issues, mainly because I don't know much about current administration of NAEP - though I do realize from reading Hazlett that administrative and bureaucratic issues have had a sharp impact on cost, design and utility of NAEP.

Cost

NAEP is without a doubt quite expensive and it is an expense borne almost exclusively by federal government coffers. There are numerous reasons for concern over this situation, but the main one is the recent general cutback in federal funds for education and research.

Wirtz and Lapointe discussed a number of different possible costsaving measures -- including very briefly the possibility of discontinuing NAEP. I won't attempt to review their suggestions here in any detail.

It is clear that there are basically two different strategies for remedying the problem of much public money going into NAEP. The first is simply to reduce the cost of NAEP, and the second is to find ways to have NAEP pay its own way with other than public funds.* Let me discuss examples with regard to each strategy.

---

*Wirtz and Lapointe's suggestions about getting states to pay more of NAEP's costs seem a bit unrealistic to me. Most state governments are under as much fiscal pressure as the federal government just now. Also, there appears to be a trend in the past 5 years or so, for states to move away from sample assessments like NAEP, toward census-type testing as in minimum competency testing.

Simply reducing costs could be accomplished in a variety of ways. Indeed it already has been, as NAEP has cut back on the assessment schedule and in recent years has largely excluded the young adult sample. However, one fairly simple way of saving money (which for reasons that escape me) has apparently not yet been tried would be to reuse released NAEP exercises. I have not seen detailed cost breakdowns for exercise development, but simply from descriptions of the process, I assume that exercise development is a fairly large expense. It is, of course, widely assumed that exercises or items disclosed in the public domain are no longer valid for operational use. Whatever the merits of this premise in general (there was considerable debate over it, for example, in hearings on "truth-in-testing" legislation), it seems to me to have little merit with respect to NAEP. The major argument typically advanced against operational use of released items is that prior familiarity with an item may invalidate it as a measure of the real skills or knowledge of interest. Subjects may have simply memorized the intended answers.

However, there are four reasons why I think this argument does not pertain to NAEP;

- First, there is such a large number of NAEP exercises that it seems very likely that very few people, if <u>anyone</u>, could familiarize themselves with (much less memorize) all or even most of the released exercises.
- Second, individuals who will be assessed lack incentive to do this. Since no direct consequence at all flow from NAEP assessemnts for <u>individuals</u> assessed, there is no reason for them a priori to familiarize themselves with or to memorize NAEP exercises.

Third, lacking incentive as a potential cause of this occurring, there is very little chance of prior familiarity leading to invalidation simply by chance. Suppose that a released NAEP item is published once in <u>every</u> newspaper in the country (very unlikely, I suspect from what I know of newspaper coverage of NAEP). Now it has been reported that only around 20% of U.S. adults read newspapers every day. Among 17-year-olds the figure is presumably lower, say 10%. If my reading habits are any guide (reading about half of what is in any one day's paper), then it would be resonable to estimate that only half of these (i.e. 5%) would read the NAEP item. Further it seems plausible that even if someone saw the NAEP item, it would be unlikely (say a 1 in 5 chance) that the person would remember the NAEP item later after a substantial period, when they happened to be in a NAEP sample. All this would mean that there would only be around a 1% chance of the sort of invalidation feared occurring with an item published in <u>every</u> newspaper in the country. Odds would presumably be <u>much</u> smaller for 9- and 13-year olds who would be less likely to read newspapers.

All this suggests that the magnitude of error deriving from use of released items would be substantally less than sampling error already implicit in the NAEP design. Obviously it would be easy to come up with numbers different than the ones advanced above, but I note that several years ago when ETS did a study of the effect of readministering the <u>same</u> forms of the SAT to the same individuals, after a period of only 3-4 months, it was estimated that the effect was fairly small -- well within the standard error for individuals if I recall correctly.

- Fourth many of the NAEP exercises are not multiple-choice but instead are open-ended or performance exercises. Invalidation resulting from prior exposure obviously is less severe with such items.

I suggest that the re-use of released exercises should definitely be investigated; first by looking at cost implications and then with some pilot work on implications of reuse for exercise validity. (Though a severe problem in the latter regard, as I discuss later, is that very little work has been done on exercise validity, apart from content validity).

The second strategy for addressing the cost problem would be not necessarily to reduce costs of NAEP, but instead to shift them from federal coffers to private ones. One idea for doing this would be to somehow make NAEP exercises commercially available for independent use. I realize that some NAEP exercises have already been used in state assessments, but this is something which ECS has done, according to Hazlett, only at cost (a strategy perhaps related to be fact that the states are the constituency of ECS). What I have in mind, however, is to sell NAEP sets of exercises (packaged in some usable form, together with interpretative materials) to schools, individuals, etc. as a means of paying for NAEP development costs. As I pointed out earlier, this would doubtless raise hackles among private testing companies, but as I argued,there are a variety of leasing arrangements which might overcome such problems. Moreover, there is clear precedent for this sort of thing with the Adult Performance Level (APL) test developed at considerable federal expense now marketed by ACT -- and the APL is much worse than NAEP exercises I have seen. Finally, quite apart from fiscal considerations, I think there are important educationl reasons for selling NAEP items, or at least for making them far more widely available.

Design and Technical Issues.

I have many observations and suggestions in this realm, but for the moment I will stick simply to points of major concern regarding validity and reliability.

As long as NAEP exercises were interpreted strictly one-by-one, without much attempt to aggregate results, I think that there were relatively few problems with regard to validity and reliability. However, now that results are regularly aggregated across objectives, sub-objectives, subject areas, etc., it seems to me that a major unresolved question concerns the validity of inferences which can be drawn from aggregated results.

NAEP validation procedures constitute what most people would call content validation. As several prominent observers (such as Messick and Cronbach) have argued in recent years, however, content validation does not necessarily constitute validity evidence at all, if we mean by this, evidence bearing on the interpretations that are warranted on the basis of test or assessment results. If we accept this line of argument, it means that far more work to be done on the construct validation of NAEP results as they are now commonly interpreted -- i.e. with results aggregated across exercises.

This is not simply an academic issue. Though I have stated previously that I thought this observation was very important, I think it even more so now that I have had a chance to inspect many of the released NAEP exercises. What I found was that the relationship between exercises and the objectives and sub-objectives to which they were assigned seemed extremely tenuous in many cases. Thus, as part of construct validation of NAEP exercise sets, I suggest that additional content validation work needs to be done. This suggestion, by the way should not be taken as a criticism of NAEP only, for these problems (that is, tenuous relationship between objectives and items, and lack of construct validity evidence) seem to me to be very common among so-called criterion-referenced or objective-referenced tests.

Construct validation is, of course, normally thought of as being carried out with respect to tests or subtests, but any of the construct validation strategies applied at the higher levels of aggregation also could be applied at the item or exercise level.

The second general technical concern I have also relates to the interpretation of aggregate NAEP results rather than interpretation of only exercise level results. My concern is that if NAEP results are to be interpreted above the exercise level (e.g. in terms of sets of exercises such as those pertaining to literal comprehension or inferential comprehension), then considerable work could usefully be done on the generalizability of NAEP results. Here I refer to generalizability theory as opposed to classical reliability theory. Without getting into a long discussion of generalizability theory (which by the way appears to be receiving considerable emphasis in the new joint test standards), let me try to explain briefly the general nature of my concern.

NAEP has long been using jack-knife estimation procedures for calculating standard errors of measurement. In several ways this practice is eminently praiseworthy. It appears to yield, for example, more appropriate estimates (implicitly taking into account multiple stages of sampling) than would procedures assuming simple random samples. By and large, I would have little quarrel with the practice of applying the jackknife procedures at exercise level (and avoiding further aggregation). The reason is that when results are reported in the form of say 50% getting exercise 8 correct when administered under XY conditions, and the exercise is presented along with the result, there is little potential for misinterpretation. This manner of interpretation makes it quite clear that results pertain to a <u>particular</u> exercise given under <u>particular</u> conditions. In the language of generalizability theory, the exercise and administrative conditions facets and variance are assumed to be fixed -- and fixed very narrowly and specifically.

However, now that NAEP, "In addition to providing results on individual items," also "reports the average performance across groups of similar items -- for the learning area as a whole, for a particular theme, objective or sub-objective, and so on" (NAEP, Reading, Writing, and Thinking report, 1981, p. xiii), it seems to me that the jackknife procedure as previously employed is no longer adequate.

There are several ways of explaining my point, but for the sake of explication let me briefly set out only one. As generalizability theory makes clear (and classical reliability theory does not) many facets or sources of variance can affect assessment results (e.g. tasks or exercises, administrative conditions, samples tested, scoring procedures, etc.). The problem with NAEP's jackknife procedure, as previously applied is that it assumes (at least as I understand it) that the only facet of error variance is the sample of individuals tested. This is not an unreasonable assumption when results are interpreted at the individual exercise level, but it seems to me potentially quite misleading when results are reported in terms of sets of items labelled such as "literal comprehension," "inferential comprehension" or "reference skills."

The reason is that other facets can and do contribute to variance in results in such areas. Exercise content and format, administrative conditions, and scoring all can contribute to error variance. Indeed, NAEP itself has pointed out that such facets can contribute substantially to variation in performance. For example, in the 1981 report Reading, Thinking, and Writing, it was written in summary that:

> The nature of a particular passage has a strong, shaping influence on the characteristics of students' responses.
>
> Item formats also have a major influence on students' performance.
>
> (p. 3)

If this is so, then the problem of interpretation for NAEP is that facets of variance concerning content samples, format samples, etc. are not fixed.*

This suggests to me that NAEP needs to become far more clear in reporting aggregate results in specifying the facets of variance over which it is and is not attempting to generalize. In the language of generalizability theory there is a need for becoming far more clear in specifying the universe of observations and conditions over which generalizations are being drawn (intentions with respect to generalizing across people are relatively clear).

Again, I strongly suspect that this is often more than a merely academic issue. It certainly appears to me (though this is an issue that I have not had time to track down thoroughly) that NAEP results may vary considerably more in terms of samples of exercises aggregated under the same label, than in terms of cycle of assessment. In other words, it may be that the content and format facets of variance are more important than the year or cohort facet of variance when results are aggregated across sets of exercises.

*Strictly speaking, NAEP does cover itself on this point, maintaining that aggregate results pertain only to "specific sets of exercises," but the manner in which sets of exercises are labelled (e.g. "inferential comprehension, as opposed to exercise set X) belies this disclaimer.

Utility

Almost every reviewer of NAEP (e.g. Greenbaum, Hazlett, GAO, Wirtz & Lapointe) has observed that NAEP results have not proven as useful as they might be. It is true, I think, as Sebring & Boruch observed of Wirtz & Lapointe that some of the conclusions regarding NAEP's utility have been rather cavelierly drawn. Nevertheless, given the substantial costs of NAEP, it obviously is worth considering ways in which NAEP could be made more useful. Here I would like to suggest two strategies: 1) developing norms for NAEP exercises, and 2) making NAEP exercises and data more readily accessible to independent investigators.

These suggestions are premised on the assumption that there are two broad types of potential use of NAEP: one by educators and evaluators using NAEP exercises and interpretative materials for their own purposes; and two by researchers and other invertigators independent of NAEP using NAEP data. In essence, of course, I am suggesting that the best strategy for enhancing the utility of NAEP may be a decentralized one -- that is not more interpretation and report peddling by NAEP itself but instead more promotion of NAEP products (i.e. exercises and performance data on the exercises).

On the first type of potential use, I have already suggested selling sets of NAEP exercises for independent use as a means of making NAEP pay more of its own way. The way NAEP is presently organized, however, I doubt that such an effort would be terribly successful. Why? Because at present there is no attractive way to make sense of the meaning of NAEP exercises. Indeed, it is quite revealing I think that NAEP's own framework for organizing exercises has changed over time -- from objectives to content by taxonomic level matrix (in at least some cases).

The most obvious way in which to make sets of NAEP exercises more attractive is to develop and publish norms for them. At first, this suggestion may seem heretical to anyone familiar with the origins of NAEP. However,

on a theoretical level, there are two considerations which indicate that developing norms for sets of NAEP exercises would not be as heretical as it first might appear. First, despite some of the rhetoric in the early days of NAEP against norm-referenced testing (e.g. Tyler arguing that selecting items in terms of difficulty and discrimination can lead to important items and objectives being overlooked), it is quite clear that NAEP has never entirely done away with normative considerations in selecting exercises. One-third were to be easy, one-third hard and one-third of middling difficulty.

Second and perhaps even more important, it is vital to distinguish between construction of test and assessment instruments and their interpretation. As is being increasingly recognized nowadays any test result, be it derived from so-called criterion- or objectives-referenced tests or from a "norm-referenced" test, can be interpreted in either criterion-referenced or norm-referenced fashion. Thus, sets of NAEP exercises could be interpreted in norm-referenced fashion (indeed they already are, as in deviation scores of regional averages from the national mean) without undercutting the distinctive character of NAEP's exercises as being developed and selected in terms of specific objective or content by cognitive level specifications.

Availability of norms for sets of NAEP exercises would greatly enhance their utility for educators and evaluators, I suspect. Moreover, developing national norms for sets of NAEP exercises could be of substantial practical interest. First, as Cooley and Lohnes have pointed out, because of its sampling procedures, NAEP has the potential for developing norms which are much more truly representative nationally than those of any of the commerical test publishers. Second, NAEP norms might shed considerable light on the debate over norm-referenced versus criterion-referenced testing. Several small-scale pieces of research have clearly shown that selection of test items in terms of prevailing standards of norm-referenced test construction can

bias the content coverage of a test. However, what has not been systematically investigated is whether or not a test such as NAEP, constructed in terms of objectives (with no screening applied in terms of item discrimination), would yield substantially different norm-referenced results (e.g. in white versus black comparisons or male versus female) than a test constructed in norm-referenced fashion. Presumably normative differences could be smaller or larger, but whichever the case the results might be of considerable practical interest.

Beyond theoretical and practical issues, making NAEP exercise sets available for local use with norms as aids for interpretation might, I think be of considerable educational interest. The reason is simply that NAEP has invested a tremendous amount of time, energy and expertise in developing exercises for educational goals which have been largely overlooked by commercial publishers (mainly for economic reasons I suspect). Making high-quality exercise sets available for such areas as music and art, which too often are neglected when it comes to assessment could be of substantial educational value.

There are, of course, many different ways in which norms could be developed for NAEP exercise sets. Some possibilities, such as grade equivalent norms, obviously should be avoided. However, there are a number of other possibilities (age and grade norms interpreted in terms of standard scores, percentiles, growth curve norms etc.) which might be considered.

The second idea I would suggest for making NAEP more useful would be to make exercises and performance data more accessible. There are many ways in which this might be done (indeed, the idea already discussed, of selling

sets of NAEP exercises is one strategy). However, as a first step in making NAEP exercises and results accessible what I would suggest is the development of a comprehensive index to NAEP exercises, surveys, reports and data tapes. In elaborating on this idea, let me first provide examples of why I think NAEP data are not very accessible, and then describe ways in which an index might make them more accessible.

First, the problem. Though I have long been interested in NAEP, only recently have I begun to review detailed information on NAEP and NAEP exercises. One thing I have done is to begin reviewing sets of released NAEP exercises. As I started doing this, I was struck by two things. First, as already mentioned, the connection between NAEP exercises and objectives seemed to me very tenuous in many cases. Second, the NAEP classification of exercises seemed to camouflage information on exercises which were of far more general interest than one would suspect by looking merely at the objective under which they were classified. This seemed particularly so in the case of open-ended exercises.

These considerations suggest to me that what would be very helpful would be a comprehensive index to NAEP exercises, surveys, reports and data tapes. Some of this information already exists I realize, for example in the identification numbers to NAEP exercises. However, it seems fairly clear to me that more thorough indexing might make NAEP exercises more accessible. Exercises might be classified for example, not only in terms of objectives on subject areas, but also in terms of vocabulary used in the exercise itself, in coding of open-ended exercises, in terms of response format (e.g. multiple choice, or open ended, written or verbal, administrative conditions etc. There are of course, many other dimensions in terms of which NAEP exercises, surveys, reports and data tapes might be coded and thereby indexed. I cannot even

begin to mention most possibilities here. Hence, let me close simply by reiterating my general point that classification of NAEP exercises in terms of subject areas and objectives seems to me quite tenuous, and that classification of NAEP materials and data more thoroughly, from several different perspectives might make NAEP more useful to people with diverse interests -- interests which often may not coincide with NAEP's objectives or content-cognitive level framework, but which might nevertheless be illuminated by the unique data set which NAEP has accumulated.