DOCUMENT RESUME

ED 262 096 TM 850 579

AUTHOR Beaton, Albert E.
TITLE NAEP Analysis Procedures and Methodology.
INSTITUTION National Assessment of Educational Progress, Princeton, NJ.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE 3 Apr 85
GRANT NIE-G-83-0011
NOTE 14p.; Paper presented at the Annual Meeting of the American Educational Research Association (69th, Chicago, IL, March 31-April 4, 1985).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Descriptive (141)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Educational Assessment; Elementary Secondary Education; Equated Scores; Item Analysis; Latent Trait Theory; *National Surveys; *Reading Tests; Research Design; *Scaling; Scoring; *Statistical Analysis; Test Construction; Trend Analysis
IDENTIFIERS *National Assessment of Educational Progress

ABSTRACT

This paper overviews technical developments in data analysis procedures for the National Assessment of Educational Progress (NAEP) reading data during 1984. The highlight of the reshaping of the NAEP data has been the scaling using item response theory (IRT). AT this point in the data analysis, an IRT-based scale appears appropriate for reading proficiency. A single dimension that spans the three grade levels (4, 8, and 11) and the three age levels (9, 13, and 17) has been located. A report of results on a scale representing a hypothetical test with known properties is in preparation. Effects of changing the administration of exercises from a tape recording to pencil-and-paper have been examined. Data from past NAEPs are being rescaled onto the new reading proficiency scales for analysis of trends. Present technology has been adapted to mesh with the new Balanced Incomplete Block (BIB) spiralling. The following activities are discussed in detail: (1) the multifaceted approach to establishing the dimensionality of the reading parameters; (2) estimation of reading parameters; (3) proficiency imputation; (4) the reading proficiency scale; and (5) trend data. (BS)

NAEP Analysis Procedures and Methodology*

Albert E. Beaton
Educational Testing Service
April 3, 1985

The past year has been a particularly exciting one for the data analysis staff of the National Assessment of Educational Progress. The new design for a new era, which was described at the last meeting of these organizations (Beaton (1984), Messick(1984)) and elsewhere (Messick, Beaton, and Lord (1983)), has been implemented; over 100,000 young people have attempted NAEP reading and writing exercises; their responses have been returned, entered into our data base, and checked; and the data analysis has started. Now, we have to show how the new design can give us more useful information about the performance of young people in the American school system.

The highlight of our reshaping the NAEP data so far has been the scaling. We have felt, and continue to feel, that summarization of the assessment data in a learning area, such as reading, into one or a few scales using item response theory (IRT) is useful, if the data are consistent with the assumptions of the theory. An important assumption of IRT is that the manifest data can be

* * *

*Paper delivered at the annual joint meeting of the American Educational Research Association and National Council for Measurement in Education in Chicago on April 3, 1985.

NAEP is sponsored by National Institute of Education Grant #NIE-G-83-0011 and administered by the Educational Testing Service.

described by a single, underlying dimension, and thus we have spent a considerable amount of time examining the dimensionality of the reading exercises. At this point in data analysis, we feel comfortable that an IRT-based scale is appropriate for reading proficiency, and have located a single dimension that spans our three grade levels (4, 8, and 11) and our three age levels (9, 13, and 17). At present, we are preparing to report results on a scale representing a hypothetical test with known properties. We have examined the effects of changing the administration of exercises from a tape recording to pencil-and-paper, and are now proceeding to rescale data from past NAEPs onto our new reading proficiency scale.

Many of the operations performed so far have required some adaptations of present technology to mesh with one of the innovative features of the new design, BIB (Balanced Incomplete Block) spiralling. BIB spiralling is a procedure by which only a small subset of the NAEP exercises is given to an individual student, but the subsets are administered in such a way that each pair of exercises is given to a nationally representative subsample of students. The use of BIB spiralling gave us a way of maintaining the assessment of a broad range of educational competencies while keeping the participation time of individual students to less than an hour, as had been done in past assessments, while also giving us the ability to compute cross-tabulations or estimate the correlation between any pair of exercises in the assessment battery. The data generated by BIB spiralling are different in form from most educational data, and so many of the algorithms for their

analysis have to be different.

It is these technical developments that will be discussed here. My intention is to give an overview of a number of the developments, and not to discuss any of them in depth; the details will be described by members of the NAEP data analysis staff in future technical reports and papers.

Before proceeding, it is important to make a general comment about assessment: we are not interested in estimating the proficiency of individual students; the NAEP assessment battery was not designed, nor is it appropriate, for that purpose. We never take a student's name outside of the school building. NAEP results are not returned to the student or to his or her teacher and thus no individual decisions can be made as a result of the assessment. Nor are results tabulated by school, so NAEP results cannot affect a school or its teachers directly. NAEP is interested in estimating the proficiency of large groups of students, and thus it is the accuracy of estimation for groups of students, not individuals, that is important. Many traditional concepts take on different meanings in this context, and, as will be shown below, we perform some operations that would be quite inappropriate for individual testing; we believe, however, that they are appropriate for the group assessment which is our goal.

Only the reading data will be discussed in this paper; we have not yet addressed many important issues for the analysis of our writing data.

Dimensionality

In A New Design for a New Era, we proposed to examine the dimensionality of a learning area using factor analysis. In hindsight, we feel that this approach is less than optimal, and so we have explored a number of additional approaches to establishing the dimensionality of the reading exercises. A full technical report on our multi-faceted approach is being prepared by Rebecca Zwick of our staff.

Our first problem in using factor analysis was computing the correlations to be analyzed. Our data contained two types of missing data, the usual type which comes from students failing to respond to items for whatever reason, and a second type which comes from the nature of BIB spiralling. The first type is not random and usually not well behaved; the second type is random, but the samples are small enough so that one cannot assume closeness to population values. However, we did compute several missing data correlation matrices of item responses. The matrix of tetrachoric correlations, even when adjusted for guessing, did not yield acceptable results; the matrix had a large number of negative eigenvalues, and some of the negative eigenvalues were large in magnitude. Several different approaches to adjustments for guessing did not help. Furthermore, when trying to establish the

dimensionality over three ages and grades, the assumption of an underlying normal distribution is unacceptable. The matrix of phi coefficients, although much better behaved in the sense that it had only a few, small negative eigenvalues, was still not good and there was no theoretical justification for using it. Finally, upon further reflection, we felt that for our exercises, with non-zero guessing parameters, the factoring of tetrachoric correlations was questionable anyway.

Thus, instead of pursuing the classical factor analytic approach only, we have tried, or are trying, several other methods:

1. Inter-block correlations corrected for attenuation. Several of the blocks of exercises were excluded from the scaling process by Dr. Frederic Lord on a priori grounds. In order to assess the differences in dimensionality between the included and excluded blocks, a number right score was computed for each block, and these number right scores were correlated and corrected for attenuation. The median corrected correlation for the included blocks was over .90, and the median correlation between included and excluded blocks was in the middle .80's.

2. Full information factor analysis (see Bock, 1984). We contracted with Prof. Darrell Bock, of the University of Chicago, to investigate the dimensionality of our data using his new method of analysis. We found that this method was too expensive for analyzing our entire data set, and so, at Dr. Bock's suggestion, we specified a subset of the data which we considered, on a priori grounds, to be most likely to generate several dimensions. The

results showed one dominant factor and several very small, but significant, other factors.

3. Rosenbaum method (1984). This method can be used to determine whether data are consistent with a model that assumes monotonicity, conditional independence, and unidimensionality. This method does not test correspondence of the data with any specific model such as the three parameter logistic model, which we used. Early results on a subset of the data did not show inconsistency with unidimensionality. More analysis using this method is expected.

4. Exploring IRT residuals. Using the IRT parameters, including proficiency estimates, this method calculates the residuals of the actual responses from those expected under the IRT model, and inspects the residuals for departures from the model. This analysis has not yet been done.

5. Several other methods, including a method of Ledyard Tucker, have been explored, but not yet applied.

The Estimation of Reading Parameters

The estimation of reading parameters takes advantage of the fact that we do not need to report scores on individual students. If we were considering individual decision-making, we would insist on administering identical or parallel tests to all students, and insist that enough items be administered to each student that his or her individual proficiency be well estimated. Since we are

interested in group estimation only, we could accept less stringent conditions, as long as the results are essentially unbiassed.

The estimation of reading parameters was done by Marilyn Wingersky using the LOGIST (1982) program.

The first step was the selection of exercises to be included in the reading proficiency scores. Some exercises were excluded for being different from what many would consider reading, such as locating places on maps or reading from tables. A Monte Carlo experiment showed us that assessment blocks with very few exercises were more harm than good in item calibration, so blocks with few exercises were excluded. All other exercises at each grade level were used, including those exercises that were used at two or three of the different age levels.

The item parameters for each age/grade combination were calibrated separately, and each set of parameters seemed to behave according to the expectations of the IRT model. Only individuals who were administered at least 17 exercises were included in the calibration.

A single set of item parameters, calibrated over all ages and grades, was then estimated in one grand run. We found that, for almost all exercises that were administered at more than one age/grade level, the item characteristic curves were essentially the same for the different ages and grades. For the few exercises that did not fit, there were obvious explanations. An example was an exercise about interpreting an allegory which the fourth graders could relate to better than the eighth graders.

Finally, a maximum likelihood estimate of each individual's reading proficiency was then computed. The results were in two metrics, the theta scale, which is an estimate of the underlying proficiency variable in standard score form, and the xi scale, which is the estimated true score on a test of 228 exercises like the exercises that were actually administered and scaled in the 1983-84 reading assessment. Proficiency scores were made for all individuals who had responded to any reading exercises, although we are using and making available only the scores of subjects who responded to 17 or more exercises. The extreme values of individual estimates were trimmed.

Proficiency Imputations

The maximum likelihood estimates of individual proficiency are problematic. Maximum likelihood estimates for individuals who responded correctly to all exercises that they were offered have estimates of plus infinity on the theta scale, and those who answer all wrong are estimated at minus infinity. Given that our subjects may have been administered only a few exercises which did not differ substantially in difficulty, we have a large number of subjects with unbelievable and unacceptable scores. The xi scale was more useful, but there were still too many extreme values. The problem was exacerbated by the fact that those subjects who were administered 30-35 exercises had fairly well-estimated proficiency scores, but the estimates were very poor for those with only a few items.

Remembering that we were not interested in individual estimation, Darrell Bock pointed us to a technology for missing data suggested by Donald Rubin (1978). Bob Mislevy (1984) is developing the application to NAEP data. The basic idea is to compute, for each subject, a posterior distribution of scores, given his or her pattern of responses, grade/age, and other concommitant information. An imputed value is then chosen at random from this distribution.

The imputed values will look like a test score, but the user must be careful. Imputed values are not intended to estimate the proficiencies of individual subjects, and should never be used or interpreted as test scores in the familiar sense of the term. Rather, the collection of imputed values over a large group of subjects can be used to estimate parameters of the distribution of proficiency in the group. We intend to put on the public use tape five imputed values per subject, each set providing as good an estimate of the population parameters as any other. We will suggest that a user run an analysis several times, using different sets of imputed values; the average of the results provides the best estimate of the parameter of interest, while the variation among them adds to the estimate of estimation error.

The imputed values are not yet ready for inclusion on the public use tape, but we expect them to be shortly. Bob Mislevy is developing the rationale and procedures for using the imputed values in data analyses.

The Reading Proficiency Scale

The next question we addressed is the form of the scale on which results are reported. Clearly, we do not want scores that would be confused with IQ, SAT, percent correct, nor grade equivalent scores. The xi scale was suggested, but this scale is dependent on the exercises which we were given by the previous grantee, and we see no reason for this set of exercises to be used as a standard for future NAEPs.

Instead, we have decided, tentatively, to report NAEP results as the score on a hypothetical test with some exemplary properties. The hypothetical test contains 500 exercises, covering the same content as the ones that we actually used. We assume that there is no guessing in this test, that is, the exercises are either open ended or have a very large number of equally attractive distractors. All exercises have the same slope, which is 1.5, the average slope of the exercises that were actually administered. We further assume that the item difficulties are equally spaced across the actual proficiency levels of our subjects, and beyond. The reader may note that the Rasch model would be appropriate for this hypothetical test, if it existed. The scores on this hypothetical test are approximately a linear function of the theta variable,

within the range of the data.

The result of this scaling is a set of scores with a theoretical range of zero to 500, but an effective range of about 100 to 400. The mean of the 4th graders is presently about 200, of the 8th graders about 250, and the mean of the 11th graders is about 300.

Trend Data

One of the other design factors of new NAEP was to collect bridge samples which administered some of the NAEP exercises using tape recorders and simple matrix sampling, as was done in past assessments. The purpose of these samples was to explore the effect of the change from tape recorder to pencil-and-paper administration and, if possible, to project the results of past NAEPs onto the new scale.

The bridge samples were collected and have been analyzed. While we have found some differences between tape and pencil-and-paper administration, we have found that the results of one method are predictable from the other, and we feel that we can map from the old data to the new. Bob Mislevy and the NAEP data analysis staff are now in the process of reanalyzing reading data from 1970, 1975, and 1980 and developing estimated scores on our hypothetical test. We intend to use these scores in the analysis of trends.

Comment

There have been, of course, a number of other developments during the past year, since we began analyzing the NAEP data. Our complex data base is in good shape, and the public use tape is available. We have produced statistical tables which give the estimated average proficiency score, and their jackknifed standard errors, for each alternate response to each background and attitude item. We have started the behavioral anchoring of the scale, and started to develop graphical methods for presentation of results. We expect the next year to be as exciting as the last.

References

Beaton, A. E., "A New Design for the National Assessment of Educational Progress," Paper delivered at the AERA/NCME meeting in New Orleans, 1984

Bock, R. D., "Contributions of empirical bayes and marginal maximum likelihood methods to the measurement of individual differences," Proceedings of the 23rd International Congress of Psychology, Acapulco, Mexico, September, 1984

Messick, S. J. "Progress toward standards as Standards for Progress: A Potential Role for the National Assessment of Educational Progress," Paper delivered at the AERA/NCME meeting in New Orleans, 1984

Messick, S. J., Beaton, A. E., and Lord F.M., A New Design for a New Era, Princeton, N. J., ETS, 1983

Mislevy, R. J., "Estimating latent Distributions," Psychometrika, September, 1984, 49, 3, 359-381

Rosenbaum, P. R., "Testing the Conditional Independence and Monotonicity Assumptions of Item Response Theory," Psychometrika, September, 1984, 49, 3, 425-435

Rubin, D. B., "Multiple Imputations in Sample Surveys," Proceedings of the Survey Research Methods Section of the American Statistical Association, 1977, 71, 538-543

Wingersky, M. S., Barton, M. A., and Lord, F. M., LOGIST Users Guide. Princeton, N. J., ETS, 1982