DOCUMENT RESUME

ED 318 786 TM 014 901

AUTHOR Beaton, Albert E.; Johnson, Eugene G.
TITLE IRT as a Way of Improving the Usefulness of Complex Data.
PUB DATE Apr 90
NOTE 15p.; Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16-20, 1990).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Achievement; Data Interpretation; *Educational Assessment; Elementary Secondary Education; *Evaluation Utilization; *Item Response Theory; *Scaling; Statistical Analysis; *Test Interpretation
IDENTIFIERS *National Assessment of Educational Progress

ABSTRACT

When the Educational Testing Service became the administrator of the National Assessment of Educational Progress (NAEP) in 1983, it introduced scales based on item response theory (IRT) as a way of presenting results of the assessment to the general public. Some properties of the scales and their uses are discussed. Initial attempts at presenting the assessment results reported the percent correct statistics for each individual item. IRT-based NAEP scales avoid the problems of average percent correct statistics and summarize complex information by reducing a large data set into a few manageable and interpretable summary statistics. The dimensionality of data sets has been studied to protect against losing important information in the summarization process. Research into scaling has demonstrated that, in most cases, the data support creating a single developmental scale. The NAEP also produces a global summary, using a weighted average of the domain scores. Scale anchoring is accomplished by describing what students at selected levels know and can do. A typical anchoring process is described. Three figures illustrate the use of NAEP scales. (SLD)

IRT AS A WAY OF IMPROVING THE USEFULNESS OF COMPLEX DATA[1]

Albert E. Beaton
Eugene G. Johnson

1. INTRODUCTION

An educational assessment is fundamentally an information system designed to report to educational practitioners, policymakers, and the general public on the status of, and changes in, the performance of students. As an information system, it must be concerned that the information produced and reported is useful to its audiences. Of course, the information presented must be accurate. But furthermore, a successful information system needs to report results in a concise and understandable way if it is to satisfy the various information needs which it attempts to serve.

The National Assessment of Educational Progress (NAEP) was designed to report what students in American schools (both public and private) know and can do. Obviously, not all of the proficiencies of students can be assessed, and so a sample of exercises in school subject areas is selected to represent the overall proficiencies of students. The selection of exercises is very important, and NAEP selects the exercises using a consensus approach. A large committee of learning area specialists, educators, and concerned citizens is formed for each subject area assessed. These learning area committees specify the objectives for the assessment in terms of goals that students should achieve during the course of their education. The objectives for a subject area are typically defined in terms of a content-by-process area matrix for which the approximate percentage of items of each type is specified by the learning area committee. In order to satisfy the objectives of the assessment and ensure that the tasks selected to measure each goal cover a range of difficulty levels, NAEP samples a large number of exercises within each school subject area that it assesses. For example, recent assessments of reading, mathematics, science, history, and civics were each based on several hundred exercises.

Simply reporting all of the data that NAEP collects to all of its audiences would be clearly inefficient and unresponsive. It would be inefficient because many different users would be required to analyze and summarize the data. It would be unresponsive since different audiences have different needs, and so information must often be tailored for them. Some NAEP users, such as educational test developers and curriculum specialists, may want detailed information for specialized purposes that were unknown and unpredictable at the time the assessment was designed, and so all of the data are made available on Public-Use Data Tapes, except for information that would compromise the anonymity of NAEP participants. Other information users, such as policy-makers and policy analysts, want more concise information that

[1]Paper delivered at the annual meeting of the American Educational Research Association in Boston, April 18, 1990.

reduces the great volume of collected data into a form that is clear and understandable and yet captures the main findings of the assessment.

There are many ways in which data may be summarized for general reporting. NAEP originally attempted to summarize the data by reporting the percentages of correct responses for individual items and later by average percentages over groups of items, but this method was found wanting for reasons to be given below. When the Educational Testing Service (ETS) became the NAEP administrator in 1983, it introduced scales based on item response theory (IRT) as a way of presenting results to the general public. These NAEP scales have now been widely accepted by many users of educational information.

In this paper, some properties of the IRT scales and their uses will be discussed. First, we will address the problems of presenting the results by the percentage of students who answered individual items correctly and also by averages of percentages-correct for groups of items. We will then discuss how the IRT captures the main features of the assessment data by scaling and how the results are reported. Some of the issues determining whether global scores or more detailed subscores will be discussed as well as the decision to make developmental scales that span different age levels. Finally, the use of scale anchoring as a way of interpreting the meaning of the scale will be presented. Further details on the characteristics of the NAEP scales can be found in the NAEP technical reports (Beaton, 1987, 1988; Johnson and Zwick, 1990).

2. ALTERNATIVE METHODS OF REPORTING

The initial attempts at presenting the assessment results reported percent correct statistics for each individual item. An (extreme) example of this type of reporting is given in Figure 1 which shows, for a single mathematics item, the percent of students, by subgroup of the population, choosing each possible response to the item. Obviously, Figure 1 provides a great deal of information (and, in fact, reports of this sort are routinely produced). However, this type of reporting quickly proved too cumbersome, and hard for the interested public (policy makers, educational researchers, interested individuals) to interpret.

For many of the constituents of NAEP, the level of detail provided by individual item level reporting is excessive, and may be counterproductive. Wirtz and Lapointe (1982) report about a rebellion of Washington D.C. parents when their children brought home mid-term report cards in the form of 4-page computer printouts indicating how the child was doing on a each item in a long list. Included in the report was an objectives check list telling, for example, whether or not the child could write a 2 digit number as a sum in which one addend is the next lower unit of ten or whether or not the child could identify initial consonant substitution in reading. Wirtz and Lapointe report that school offices were deluged with requests about how to interpret all this information.

The problem with an item by item approach of reporting is that it ignores overarching similarities in trends and group comparisons that are common across items -- exactly what an assessment is supposed to identify. Bock, Mislevy and Woodson (1982) distinguish between the "fixed-item" approach of reporting and the "random-item" concept. The fixed item approach assumes that each individual item is of primary interest of itself - the model for this approach is survey research where each response (opinions on some issue for example) is of unique interest. Educational items generally do not share this unique importance. Rather, the items are viewed as random representatives of a conceptually infinite pool of items within the same domain and of the same type. In this random item concept, a set of items is taken to represent the domain of interest.

Having moved away from primary emphasis on individual items to domains represented by sets of items, there is the need to have some measure of the achievement within each domain. An obvious index is the average percent correct across all presented items within the domain of interest. As noted by Bock, Mislevy and Woodson, the averaging tends to cancel out the effects of peculiarities in item writing which can affect item difficulty in unpredictable ways and produces the central tendency of the distribution of correct responses across the presented items as the measure of the overall achievement within the domain.

As noted by Wirtz and Lapointe, the process of aggregation for the reporting of NAEP results was generally applauded. Media coverage of assessment results dramatically increased when aggregate results were reported. This was at least in part because the aggregate reports provide the broad picture that the media and the public alike seem to consider more interesting and informative than individual item-by-item results. However, there are a number of significant problems with average percent correct scores.

First, the interpretation of average percent correct results depends on the selection of items; the selection of easy or difficult items could make student performance look good or bad - to a public accustomed to a passing score. Second, the average percent correct metric is obviously tied to the particular items going into the average. This means that age-to-age and year-to-year comparisons require the same exercises. The consequence of this is that measurement of trends in achievement based on an average percent correct metric is limited to either:

1) considering trends on the small proportion of items common across years (a proportion which is continually diminishing because of the need to release items and the fact that some items become outdated)

or

2) pairwise comparisons from one year to the next based on the shared items. This can lead to different and incomparable percent-correct scales for different comparisons. As an example, trends in science achievement for the 1970, 1973 and 1977 assessments had to be expressed in terms of the relative change from one year to the next.

Finally, it is difficult to speak in terms of the distribution of proficiencies in the population when the measure used is the average percent correct. As noted in the report of the NAEP planning project under the auspices of the Council of Chief State School Officers (1988), reporting an average score for a population provides, by itself, very little information to the public, policy-makers or practitioners. Averages from two or more assessments can indicate trends in central tendency, but provide no information about how the performance of the students is distributed. What is needed is information about trends in the distribution of student achievement, that is changes in the proportions of students with subject area abilities at or above specified levels.

The IRT-based NAEP scales provide a meaningful description of achievement while avoiding the problems of average percent correct statistics. The NAEP scales allow all students to be placed on a common scale even though none of the respondents take all of the items within the pool. This is important because it allows us to measure trend in terms of a consistent metric even though the item pool evolves over time. Furthermore, we can estimate the distribution of skills among students and provide meaningful interpretations about score levels in terms of predicted performance on specific exercises and in terms of educationally relevant tasks.

3. THE NAEP SCALES

Before proceeding, it is important to note that, with real data, no summary is able to capture all of the information that is available in the data. The most common statistic for summarizing data on a single variable, in education or other research areas, is the arithmetic average or mean, which captures only a small part of the information that is available in an entire data set. The mean is the single number that reproduces the original data most accurately in the least squares sense. Using two statistics, the mean and the standard deviation, gives more information about the available data, but not all, unless the distribution of data can be described by a known distribution such as the normal. Percentiles give even more information, but still do not completely describe the original data when the sample size is large. Summarizing, therefore, always loses some of the detailed information that is available in the full data set.

Although scaling educational data inevitably loses some information about the performance of examinees, it is nonetheless an efficient way of summarizing the complex assessment data. In most sets of data resulting from educational tests, there are regularities in the data. Response patterns tend towards a triangular shape in a subject-by-items data matrix that is sorted both by the number of correct responses and item difficulty (see Figure 2). Some examinees answer a large number of items correctly, some answer only a few. Some items are "easy" in that they are answered correctly by most examinees, others are "hard" in that only a few persons answer them correctly. And in most data sets, there is an interaction between the examinees and the items, that is, the examinees who answer the difficult items correctly tend to answer the easy ones correctly too, but answering the easy items correctly

does not imply answering the more difficult items. This is the regularity in the data that IRT scaling attempts to encapsulate.

Scaling is a process by which the data are reduced to a few statistics that summarize a large data set. One or more statistics may be defined to represent the proficiency of examinees on one or more dimensions of proficiency. One or more statistics may be defined to represent the characteristics of the different items. An example of a simple scale is the number-right score which represents the proficiency of examinees and is particularly useful when all examinees have responded to the same set of items. The proportion of examinees passing an item is a statistic that describes a property of an item.

In most applications of item response theory, a single scale score represents an examinee's proficiency, and this score can be used to estimate an examinee's response to any item in the examination. The relationship between an examinee's response to an item and his or her scale score is approximated by a non-linear function. The non-linear function, which is different for each item, is characterized by an item statistic or statistics defined for each item. These item statistics are estimates of item parameters defined over a population of examinees. The item statistics may estimate one parameter, as in the Rasch model, or three parameters, as in the IRT model used by NAEP. The three parameter logistic model that is used in NAEP has been found to fit many data sets reasonably well.

No real data set is perfectly ordered and thus any scaling procedure that reduces the test results into a few summary statistics must lose some information. Examinees who are generally able to answer items may have occasional gaps in their knowledge. Especially when multiple choice items are used, poorly performing students may occasionally respond correctly to difficult items, even if by chance. As we shall see below, these departures from the expected regularity should be studied separately to help understand what the scale values encompass and what they do not. In most cases, the scale values will encompass most of the general information about item responses that is in the data.

This is the real function of IRT in an assessment: to summarize complex information by reducing a large data set into a few manageable and interpretable summary statistics. From the individual scores, we can estimate--and to some degree reproduce--the individual's actual responses to items. Individuals with high scores tend to get many items correct; individuals with low scores do not. If different items were randomly assigned to examinees, then we can also estimate how an examinee would have performed on items that he or she was not administered. The score is, therefore, a simple summary of how the individual did on the many items in the assessment.

IRT scaling is not the only way to summarize data. When writing was assessed in 1984, several attempts were made to adapt IRT technology to the nonbinary writing exercise responses using the models proposed by Masters (1982). However, it was found that these models did not provide acceptable results for the NAEP data and so an alternate method of scaling called the

Average Response Method (ARM) of scaling was developed and used. The ARM method is fully described in Beaton and Johnson (1990).

The NAEP reporting scale

In addition to summarizing the data, NAEP must report its results. An IRT scale is indeterminate; any linear transformation of the scale will reproduce the item responses equally well. By default, most computer programs assign the average scale score to be zero and the standard deviation to be one with the result that about half of the scale scores are usually negative. Such an arbitrary reporting system would not be appropriate for public dissemination.

NAEP reports its results as number-right scores on a hypothetical 500 item test. Scores on this hypothetical test typically range between about 100 to 400. The test has certain idealized properties: the item difficulties of this hypothetical test are distributed evenly across the range of observed performance and somewhat beyond, all items have the same discriminatory power, and there is no guessing. If such a test could be built, it would fit the Rasch model precisely.

Using this hypothetical test, NAEP results can be interpreted as test scores on an idealized test. There are no negative scores. Although the number of items in the test is arbitrary, it was chosen to minimize confusion with IQ scores, SAT scores, grade equivalents, and other well-known test metrics.

4. MULTIVARIABLE SCALING

As mentioned above, the objectives of national assessment instruments are developed by an extensive consensus process that reviews the curriculum objectives of a subject area. In some cases, the assessment is designed to assess different types of proficiencies which may be taught at different levels or in different subject-area courses. In the 1986 mathematics and science assessments, the objectives committee decided that different sub-domains in each area were important enough to be assessed and reported separately. In mathematics, the sub-domains were (1) Knowledge and skills in numbers and operations, (2) Higher level applications of numbers and operations, (3) Measurement, (4) Geometry, and (5) Algebra. In science, the sub-domains were (1) Life Sciences, (2) Physics, (3) Chemistry, (4) Earth and Space Sciences, and (5) the Nature of Science. The item response theory approach to scaling was adapted to fulfill the requirement of separate reporting by sub-domain.

In the cases of mathematics and science, student exercises were developed to probe each sub-domain. The assessment items, therefore, could be classified by sub-domain. However, the amount of individual student time for NAEP is kept to about an hour and thus accurate measurement of students in each sub-domain was impossible. Simply applying existing IRT technology would have resulted in poor estimates of student performance in the several sub-

domains. The concepts of plausible values and conditioning (see, e.g., Mislevy, 1990; Johnson, 1989) were applied to the NAEP mathematics and science assessments to generate, under certain reasonable assumptions, consistent estimators of attributes of the distributions of student performance in the sub-domains.

5. DIMENSIONALITY

As previously mentioned, summarizations necessarily lose some of the information that is available in the original data. The IRT scaling used in NAEP is intended to summarize the previously defined subject areas or sub-domains of those subject areas. In so doing, the IRT process assumes that there is a regularity in the data in the subject area or domain that is scaled--but what if there is not? Is there a better way to summarize the data, perhaps using several different scales? Can the data be summarized at all?

NAEP has addressed the question of what to summarize and report through the study of the dimensionality of the data within a subject area or domain. After the 1984 assessment of reading, the dimensionality of the reading data was extensively studied, and it was concluded that reporting several reading sub-scales would not add important information to reporting one reading scale. Various studies of the several scales in mathematics and science have shown that the different subscales do provide additional information that a single scale would not--for example, gender differences in mathematics and science performance--and no reason has been found to subdivide further the domains that have been reported. Studying the dimensionality of data sets protects against losing important information in the summarization process.

6. DEVELOPMENTAL SCALES

When summarizing, a natural question arises as to what populations should be summarized over. NAEP has traditionally assessed 9-, 13-, and 17-year olds and has recently added overlapping samples of fourth, eighth, and twelfth grade students. Should the scales be separately defined at each age or grade level or can one scale span all three age and grade levels? An advantage of separate scales is that the information may be targeted to specific age levels. An advantage of a single scale is that performance and changes in performance of each grade level can be viewed in the context of student growth between age and grade levels.

NAEP has approached this question empirically and found that in most cases the data support using a single, developmental scale. In designing the NAEP instruments, NAEP has included common items at adjacent age levels, that is, the fourth and eighth grade assessments include some common items as do the eighth and the twelfth grade assessments. These items allow us to compare

the performance of the older students with that of the younger students. From the perspective of data summarization, the question is whether or not a single item response function fits adjacent age levels equally well; that is, does assigning a common scale score to students at adjacent age levels seriously affect the .curacy of the prediction of their responses on common items?

Before a developmental scale is produced, the empirical estimates of item characteristic functions are obtained for each item for each age level that was presented the item. These empirical item characteristic functions are calculated without assuming any functional form thereby allowing checks of the assumptions of item parameter independence and unidimensionality. In a small number of cases (less than 4%) the empirical item characteristic functions differ between ages. In such cases, the item is treated as a separate item at each age.

The research into scaling has demonstrated that, in most cases, the data do support creating developmental scales. Using this feature, we are able to display trends in student proficiency in reference to the differences between age levels. It has also been possible to compare the proficiencies of students at one age level to those of another. For most items (excluding those mentioned above), we can say that students with a given scale score have approximately the same probability of answering the common items correctly regardless of their ages or grades. If the common items can be assumed to be from the same item population as the other items, then we can estimate how students at one age level would perform if given items from another.

7. GLOBAL SCORES

As mentioned above, different levels of summarization are appropriate for different audiences. For very detailed analyses, the basic data are made available so that any data analyst may summarize the data in any way he or she thinks appropriate. NAEP has chosen to report mainly at the level of subject area and domain, and uses dimensionality analyses to investigate the information that is not included in the summary. However, when reporting is done by domains within a subject area, many policy-makers want a more global summary, and so NAEP produces an overall summary also.

To compute an overall summary, NAEP uses a weighted average of the domain scores. When the domains are defined, the learning area committee of subject matter specialists, educators, and others not only specifies the assessment objectives but also gives each objective a weight signifying its importance. These weights are used to determine the number of items in the assessment that will be used to measure each objective. These weights are also used in computing the overall summary of performance.

It is clear that the overall summary does not contain all of the information in the more detailed domain scores--if it did, there would be no point to estimating performance in the separate domains. The domains themselves do not contain all of the information that more detailed summaries

might have, and so on. And yet, from some perspectives the overall summary is adequate and useful for some information purposes.

8. ANCHORING

Scale anchoring is a way of attaching meaning to a scale. Traditionally, meaning has been attached to educational scales by norm-referencing, that is, by comparing students at a particular scale level to other students. In contrast, the NAEP scale anchoring is accomplished by describing what students at selected levels know and can do. This is the primary purpose of NAEP.

The anchoring process is straightforward. There are several ways to anchor a scale, and the conceptually simplest will be described here. Several scale levels are selected--they should be far apart to be noticeably different but not so far apart as to be trivial. The students are then sorted by scale score, and students at or near each level are grouped together. For the group at the lowest scale score level, what they could know and can do is defined by the items that a vast majority of the students answered correctly. At the higher score level, the question is: what is it that students at this level know and can do that students at the next lower level cannot. The answer is defined by the items that a vast majority of students at this level answered correctly but a majority at the next lower level answered incorrectly. The assessment items are, therefore, grouped by the levels between which they discriminate. Many items do not discriminate between any pair of scale points.

Figure 3 shows a graphical representation of the statistical anchoring process. Three items are displayed, identified by the labels "A", "B", and "C". Six anchoring levels are identified, corresponding to scale values of 100, 150, 200, 250, 300 and 350. An item will anchor at a given level if: (1) 65 to 80 percent of students attaining that level can answer the item, (2) the probability of success on the item for students at the next lower level is less than 50 percent, and (3) the difference in the probabilities of success between the two levels is at least 30 percent points. In Figure 3, Item "A" anchors at the 250 level since the probability of correct response for students with proficiencies around 250 is 80% while the probability of success for students at the next lower level (200) is 40%. Item "B" anchors at the 300 level since there is a steep rise in the probability of success between 250 and 300 and since the probabilities of success at the two levels satisfy the threshold values. Item "C" does not anchor at any level because the discrimination between adjacent levels is not sufficiently sharp.

A committee of subject matter experts, educators, and others is then assembled to review the items and, using their knowledge of the subject matter and student performance, try to generalize from the items to more general constructs. Several sample items are selected to illustrate the construct. The constructs are then described verbally and sent out for general review by professionals in education.

SUMMARY

The purpose of NAEP is to assess and report what students in American schools know and can do. Over the years, NAEP has pursued this purpose in a number of ways. The introduction of items response theory into NAEP, and several innovations in IRT, have resulted in scales that summarize the main findings in the vast set of student performance demonstrations that NAEP collects. These scales are not only useful to the educational research community but are also useful in communicating information about the status of education in America to the general public. NAEP not only summarizes the data but also makes all of its data available to the research community for detailed analyses or alternate summarizations.

REFERENCES

Beaton, A.E. (1987). *Implementing the new design: The NAEP 1983-84 technical report*. (No. 15-TR-20). Princeton, NJ: Educational Testing Service.

Beaton, A.E. (1988). *Expanding the new design: The NAEP 1985-86 technical report*. (No. 17-TR-20). Princeton, NJ: Educational Testing Service.

Beaton, A.E. & Johnson, E.G. (1990). The average response method of scaling. *Journal of Educational Statistics*, in press.

Bock, R.D., Mislevy, R.J., & Woodson, C.E.M. (1982). The next stage in educational assessment. *Educational Researcher, 11*, 4-11, 16.

Council of Chief State School Officers. (1988, March). *On reporting student achievement at the state level by the National Assessment of Educational Progress*. (Report of the National Assessment Planning Project, Wilmer S. Cody, director). Washington, DC: Author.

Johnson, E.G. (1989). Considerations and techniques for the analysis of NAEP data. *Journal of Educational Statistics, 14, No.4*, in press.

Johnson, E.G., & Zwick, R. (1990). *The NAEP 1988 technical report*. Princeton, NJ: Educational Testing Service.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Mislevy, R.J. (1990). Randomization-based inference about latent variables from complex samples. *Psychometrika*, in press.

Wirtz, W., & Lapointe, A. (1982). *Measuring the quality of education: A report on assessing educational progress*. Washington, DC: Author

Figure 1: Example of a complete item-level report
(Source: Mathematics Report 04-MA-20, 1972-73 Assessment, National Assessment of Educational Progress)

RP02 P55002-34 P55002-34 RP02

	NATION	REGION				SEX		RACE		PARENTAL EDUCATION				SIZE-AND-TYPE OF COMMUNITY						
		SOUTH	EAST	WEST	CENTRAL	NORTH EAST	MALE	FEMALE	BLACK	WHITE	NO HS	SOME HS	GRAD HS	POST HS	LOW	EXTREME RURAL	SMALL PLACES	MEDIUM CITY	MAIN BIG CITY	URBAN FRINGE	HIGH METRO
EXERCISE: RP02																					
A car can be bought for cash for $2,850 or on credit with a down payment of $400 and $80 a month for 3 years. How much MORE is buying on credit than paying with cash?																					
RESPONSE: 00 AGE: 17	TEXT: No Response.																				
DELTA-P	1.9	2.5	-0.9	0.1	-1.3	0.5	-0.4	5.2*	-1.1*	0.7	3.4	-0.6	-1.1*	0.0	-1.2	-1.3*	-1.1	15.6*	-1.0	-1.9*
STD ERR OF DP		1.8	0.7	1.0	0.7	0.5	0.4	2.1	0.4	1.2	2.3	0.5	0.4	1.2	0.9	0.6	0.7	7.5	0.9	0.6
STD ERR OF P	0.6	2.3	0.4	1.2	0.5	1.0	0.4	2.6	0.4	1.0	2.8	0.4	0.4	1.0	0.7	0.2	0.5	8.1	0.5	0.0
AGE: Ad																					
DELTA-P	0.2	0.2	-0.1	-0.1	0.0	0.0	-0.0	0.1	0.0	-0.1	-0.2*	0.0	0.1	-0.2*	0.0	-0.1	-0.2*	0.2	-0.2*	0.3
STD ERR OF DP		0.2	0.1	0.2	0.2	0.1	0.1	0.3	0.0	0.2	0.1	0.2	0.2	0.1	0.4	0.1	0.1	0.4	0.1	0.4
STD ERR OF P	0.1	0.2	0.1	0.2	0.3	0.2	0.1	0.3	0.1	0.1	0.0	0.2	0.3	0.0	0.4	0.1	0.0	0.4	0.0	0.5
RESPONSE: 01* AGE: 17	TEXT: All Acceptable Responses. (Only categories 10 and 11 were considered to be the acceptable responses.)																				
DELTA-P	53.8	-8.8*	-2.0	2.5	6.1*	-0.4	0.3	-31.0*	6.8*	-17.8*	-11.7*	-2.1	12.4*	-16.3*	-0.4	0.3	2.0	-13.3	4.5	18.5*
STD ERR OF DP		3.5	2.7	2.6	2.4	1.4	1.2	2.7	0.9	4.5	3.1	1.8	1.4	4.6	4.1	2.0	4.0	7.0	2.8	4.5
P-VALUE	55.8	46.9	53.7	58.2	61.8	55.4	56.1	24.8	62.5	37.9	44.0	53.7	68.2	39.5	51.3	56.1	57.8	42.5	60.3	74.3
STD ERR OF P	1.6	4.0	2.9	3.1	2.8	2.1	2.1	2.8	1.4	5.0	3.8	2.1	1.8	5.0	4.1	2.5	4.0	7.4	2.5	4.8
AGE: Ad																					
DELTA-P	67.7	-0.6	1.5	0.0	-1.2	6.6*	-6.2*	-30.4*	5.7*	-6.9*	-5.7	3.2	12.5*	-20.5*	3.0	1.0	-0.7	-3.0	1.7	10.8*
STD ERR OF DP		2.8	3.1	2.3	2.7	1.4	1.3	3.7	0.8	2.6	3.5	2.0	2.2	4.2	3.9	1.9	4.4	4.2	4.1	2.8
P-VALUE	67.7	67.1	69.2	67.7	64.5	74.3	61.5	29.3	73.4	60.8	62.0	70.9	80.2	47.2	70.7	68.7	67.0	68.6	69.1	78.5
STD ERR OF P	1.6	3.0	3.8	2.5	3.1	1.9	2.2	3.6	1.4	3.1	3.9	2.3	2.3	4.5	3.8	2.5	4.7	4.2	4.6	2.8
RESPONSE: 10 AGE: 17	TEXT: $430: 430 dollars																				
DELTA-P	51.7	-9.0*	-2.0	3.9	5.6*	-1.2	1.0	-32.8*	7.1*	-16.2*	-13.4*	-2.4	13.0*	-16.2*	-3.7	0.1	0.1	-10.3	4.3	19.3*
STD ERR OF DP		3.7	2.7	2.8	2.7	1.4	1.2	2.8	0.9	4.2	3.0	1.8	1.4	4.7	4.4	2.2	4.2	7.1	2.6	4.9
STD ERR OF P	1.7	4.2	2.7	3.4	3.3	2.2	2.1	2.7	1.6	4.6	3.6	2.2	2.0	5.0	4.4	2.8	4.2	7.5	2.6	5.2
AGE: Ad																					
DELTA-P	57.2	0.8	-1.4	-0.5	1.4	4.5*	-4.1*	-38.8*	5.4*	-9.6*	-7.1*	7.0*	10.9*	-19.8*	2.7	1.1	0.2	1.6	-3.4	11.3*
STD ERR OF DP		2.7	3.6	2.0	2.6	1.4	1.3	3.1	0.7	2.4	3.5	2.0	2.0	3.5	5.0	1.8	4.1	4.0	4.9	2.9
STD ERR OF P	1.6	3.0	4.6	1.9	2.8	2.2	2.0	3.1	1.5	2.9	3.7	2.4	2.4	3.6	5.2	2.3	4.4	4.1	5.4	3.3
RESPONSE: 11 AGE: 17	TEXT: 430 with no sign or wrong sign																				
DELTA-P	4.0	0.2	-0.1	-0.5	0.5	0.8	-0.7	1.8	-0.3	-1.7	1.7	0.4	-0.6	0.1	-0.8	0.2	1.9	-3.0*	0.2	-0.8
STD ERR OF DP		0.8	0.8	1.0	0.7	0.5	0.4	1.3	0.3	1.3	1.0	0.7	0.5	1.4	1.0	0.9	1.8	0.6	1.3	1.2
STD ERR OF P	0.5	0.8	0.9	1.3	0.8	0.8	0.6	1.4	0.6	1.1	1.1	0.8	0.7	1.4	0.9	1.2	1.7	0.4	1.4	1.2
AGE: Ad																					
DELTA-P	10.5	-1.4	2.9*	0.5	-2.5	2.2*	-2.0*	-3.7*	0.3	2.7*	1.4	-3.8*	1.6	-0.7	0.4	-0.0	-0.9	-4.7*	5.1*	-0.5
STD ERR OF DP		1.5	1.4	1.6	1.5	0.8	0.7	1.8	0.3	1.3	1.7	1.0	1.2	2.2	2.0	1.1	1.6	2.0	2.4	2.6
STD ERR OF P	0.9	1.7	1.5	2.0	1.7	1.1	1.2	1.8	1.0	1.8	2.0	1.1	1.4	2.1	2.0	1.7	1.7	2.0	2.5	2.7
RESPONSE: 12 AGE: 17	TEXT: Correct process with addition or multiplication error; any decimal of 430																				
DELTA-P	6.7	-0.2	.1	0.4	-0.4	0.3	-0.3	-3.6*	0.8*	-0.2	-2.6	1.4	-0.1	-2.6*	-0.0	0.1	0.1	1.2	2.6	-2.2
STD ERR OF DP		1.0	0.9	1.0	0.8	0.7	0.6	0.9	0.3	1.8	1.4	0.9	0.6	1.1	2.0	0.8	1.3	1.7	1.4	1.3
STD ERR OF P	0.5	1.1	1.0	1.3	0.9	0.9	0.8	0.9	0.7	1.9	1.3	1.1	0.8	1.2	2.1	1.0	1.3	1.8	1.5	1.3
AGE: Ad																					
DELTA-P	7.0	0.7	-0.6	0.7	-0.8	-0.9	0.8	2.7	-0.0	-0.5	4.1*	-0.3	-1.7	1.4	0.2	-0.3	3.3	-0.6	0.2	-3.1*
STD ERR OF DP		1.3	1.6	1.1	1.1	0.8	0.8	2.0	0.3	1.0	1.9	1.1	0.9	2.5	2.5	0.9	3.5	1.7	1.7	1.3
STD ERR OF P	0.8	1.4	2.1	1.3	1.2	1.2	1.0	2.1	0.9	1.4	2.2	1.3	1.0	2.6	2.6	1.2	3.9	1.6	1.7	1.3
RESPONSE: 13 AGE: 17	TEXT: 3280 - 2850 with subtraction error																				
DELTA-P	0.9	-0.0	0.1	0.3	-0.4	0.4	-0.3	-0.7*	0.1*	-0.7*	0.9	0.2	-0.1	-0.7*	1.4	-0.3	-0.4	2.0	0.6	-0.9*
STD ERR OF DP		0.4	0.4	0.4	0.3	0.3	0.2	0.3	0.1	0.3	1.0	0.3	0.2	0.2	1.1	0.2	0.5	1.3	0.8	0.2
STD ERR OF P	0.2	0.5	0.5	0.5	0.3	0.4	0.3	0.1	0.3	0.2	1.1	0.4	0.1	0.1	1.2	0.2	0.5	1.4	0.9	0.0

Figure 2: Response Patterns for an Ideal Test

Subject-by-item data matrix sorted by number of correct responses and by item difficulty

	Items													
	1	2	3	4	5	6	7	8	9	10	11	12	...	m
High Scorers	1	1	1	1	1	1	1	1	1	1	1	1		1
	1	1	1	1	1	1	1	1	1	1	1	1		0
	1	1	1	1	1	1	1	1	1	1	1	1		0
	1	1	1	1	1	1	1	1	1	1	1	0		0
	1	1	1	1	1	1	1	1	1	1	0	0		0
	1	1	1	1	1	1	1	1	1	0	0	0		0
	1	1	1	1	1	1	1	1	0	0	0	0		0
	1	1	1	1	1	1	1	0	0	0	0	0		0
	1	1	1	1	1	1	0	0	0	0	0	0		0
	1	1	1	1	1	0	0	0	0	0	0	0		0
	1	1	1	1	0	0	0	0	0	0	0	0		0
	1	1	1	0	0	0	0	0	0	0	0	0		0
	1	1	0	0	0	0	0	0	0	0	0	0		0
Low Scorers	1	0	0	0	0	0	0	0	0	0	0	0		0
	0	0	0	0	0	0	0	0	0	0	0	0		0

Figure 3: Three example items for Scale Anchoring
Item "A" Anchors at 250
Item "B" Anchors at 300
Item "C" Does Not Anchor