Contributions of Background Questions to Improving the Precision of NAEP Results[1]

Mark D. Reckase
Michigan State University

Introduction

The National Assessment of Educational Progress (NAEP) is a large scale survey of the functioning of the educational system in the United States. NAEP has two major components: cognitive tests that measure the achievement of students on well defined content, and non-cognitive survey items that elicit information from students, teachers, and school administrators about demographics and the educational process. The major focus of this paper is on the contribution of the non-cognitive component, typically called "background questions," to the estimation of the distribution of student achievement.

The background questions have three functions within NAEP. First, they are used to define subgroups of the examinee population for reporting purposes. For example, NAEP results are reported by gender, race/ethnicity, parents' highest level of education, type of school, participation in Title I, and eligibility for free/reduced-price lunch. In order to report results for subgroups defined by these demographic categories, the necessary background information must be collected.

The second function for the background questions is to support research on the factors that affect NAEP scores. The background questions related to this function describe educational activities that take place at home and in school. For example, *The 1998 NAEP Reading Report Card* (NCES, 1999) reported on the effects of television viewing, daily reading habits, classroom reading and writing assignments, and discussion of schoolwork at home on the achievement of students. Information on these activities had to be collected as part of NAEP for these relationships to be investigated.

The third function of the background questions is to improve the estimation of the students' proficiency distribution on the cognitive component of the NAEP. The NAEP cognitive assessments of student proficiency are designed to perform a very challenging task. The goal is to assess what students know and can do in very broad subject matter areas in a way that is consistent with the varying educational systems across the country and curriculum standards documents, but to limit the amount of intrusion into the schools so that students will be motivated to perform well and so that participation rates will be high.

[1] Paper prepared under contract to the National Assessment Governing Board, September 2002.

Achieving this goal requires an assessment tool with a substantial number of items sampled from a broad domain, but one that can be administered in the limited time of about one hour. To meet these seemingly contradictory conditions, NAEP uses a matrix sampling design for the test. This means that each student takes only a relatively small sample of the items that assess the domain. This design does not allow accurate estimation of individual student performance, but it does allow the estimation of characteristics of the distribution of performance for the target student population and selected subpopulations.

The set of items taken by each student is carefully specified so that the individual student results can be aggregated to estimate the distribution of performance of all the students in the population on the full domain of items. The estimation of this distribution is a statistically challenging activity. Part of the estimation process uses the information from the background questions to augment the information contained in the cognitive items. The main function of this paper is to discuss how this process works, the effects of the background questions on the estimates of the proficiency distributions, and how the selection of the background questions might be changed to improve the process.

The NAEP Estimation Process

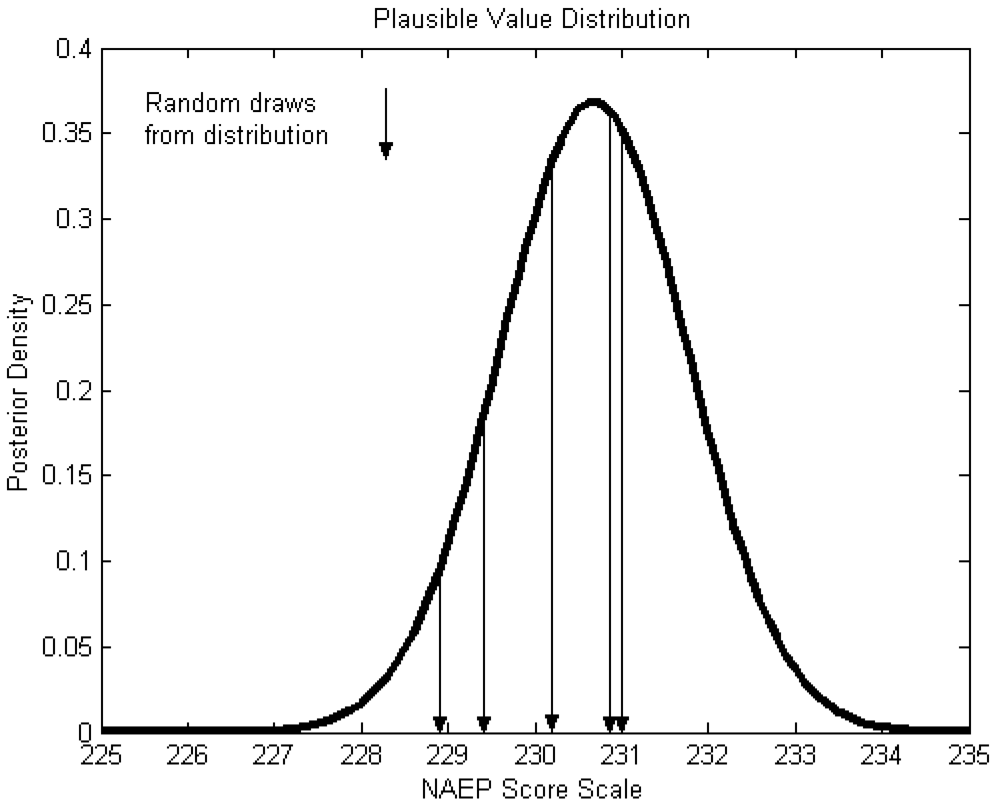
The design of the assessment requires that the assessment tasks (items) be divided into sets with relatively short administration times called "blocks." Each student is administered a carefully selected set of blocks of items. A test booklet for a student usually has three blocks of cognitive items and two sections of background and attitude questions, the non-cognitive items. Blocks are assigned to booklets so that booklets will have blocks in common with other booklets. The blocks are also arranged so that they will occur in each position in the booklet – that is, first, second, or third. This approach balances out possible position effects such as fatigue at the end of a booklet, or anxiety at the beginning of a test.

The following table gives a simplified example of the assignments of blocks to booklets and students (after Johnson, 1992). This example has seven blocks overall and three blocks per booklet. Each row of the table corresponds to a booklet. Note that Block 1 is in the first position in Booklet 1, the second position in Booklet 7, and the third position in Booklet 5. The block is also not adjacent to the same blocks in any two booklets. Booklets are randomly assigned to student samples. The full design is sometimes called balanced incomplete block (BIB) spiraling because the positions of the blocks are balanced in location, not all possible combinations of blocks are used (the set is incomplete), and the process of distributing booklets to students to get equivalent samples is called spiraling.

Booklet/ Student Sample	Blocks Contained in a Booklet		
1	1	2	4
2	2	3	5
3	3	4	6
4	4	5	7
5	5	6	1
6	6	7	2
7	7	1	3

The fact that no student takes all of the items and no booklet is administered to all of the students means that there is no way to directly get an accurate proficiency estimate based on all of the items for each student. Instead, the information from the items students did take and the information in the background questions are used to estimate a distribution of possible proficiency estimates that the student might have gotten get if he or she took all of the items. This distribution is sometimes called a plausible value distribution because it tells what a student will likely do (what is plausible).

The plausible value distribution is not used directly to estimate the population distribution. Instead, five values are randomly selected from the distribution. These randomly selected values are called “plausible values.” An example of a plausible value distribution for a student on NAEP Reading is given below. The arrows in the distribution show five randomly selected plausible values from the distribution.



The plausible value distribution is estimated from two sources of information. First, the background questions are used to predict the distribution of likely performance for a student. This is called a “prior” distribution because it comes prior to the consideration of responses to the cognitive items. The information from the cognitive item responses is combined with the prior distribution to form a “posterior” distribution. This posterior distribution is the plausible value distribution.

The logic of the estimation process is essentially as follows:

1. If we know nothing about an examinee, our best guess of his or her proficiency level is the average value of the population proficiency distribution. However, there is a lot of error in that estimate.
2. The background variables can be used to predict the score for the examinee based on the relationship between the background variables and the proficiency trait. If the relationship is strong, the predicted values will have little error. If the relationship is weak, the error in the predicted values will be about the same as guessing the average for the full distribution.

3. The actual responses to the cognitive items sampled for the examinee are used to refine the prediction made from the background variables. If the set of cognitive items provide a lot of information about the examinees proficiency, that information overrides the prediction from the background variables. However, if there is little information in the responses to the items, for example if the items are very difficulty for the examinee, then the predictions from the background variables will have a notable effect.
4. The posterior distribution provides the combined information from the responses to the cognitive items and the background variables for an examinee. Five plausible values from that distribution are sampled and aggregated over examinees to estimate the full distribution of performance for all of the examinees on the trait defined by all of the items.

Although these four steps provide the basic framework for the NAEP estimation methodology, in many cases the process is complicated by other factors in the design and implementation of NAEP. One such factor is the fact that NAEP tests contain multiple content areas that sometimes define multiple scales. Rather than estimating a single score, the procedure must estimate multiple scores, for example five subscales in mathematics, and then combine them to form a composite score. Items for the multiple scales must be included in the blocks used to construct the test booklets for the plausible values for those scales to be estimated for an individual. Sometimes, however, it is not possible to include all of the content in a single booklet. That is, all booklets do not have the same content distribution because of practical constraints of passage size or length of time needed to do an open-ended performance task.

A second complication is that background questions often indicate membership in categories such as region of the country rather than a quantitative variable such as years of parent's education. Categorical variables do not work well with the statistical procedures used to predict the prior distribution. They must be converted into 0/1 variables for analysis. The background questions or the coded variables may also be highly intercorrelated yielding redundant information. After coding into 0s and 1s, the number of individual variables corresponding to the background questions can be over 1000 for a particular grade level and content area.

Each of these variables provides relatively little information in a discrete form. To improve the functioning of these variables, statistical procedures are used to form composites of background variables (using principal components analysis) called conditioning variables. This procedure first finds the composite of variables that accounts for the greatest amount of variation in the variables. That composite is Component 1. Next it finds another composite of variables from what is left in the relationships after the amount explained by Component 1 is removed. This second composite is Component 2 and it is uncorrelated to Component 1. That is, the two components contain unique, unduplicated information. Then the relationships explained by Component 2 are removed and another composite is formed that explains the greatest amount of variation in what is left, Component 3. Each subsequent component explains less of the variation than the previous component. Component 1 explains the most. To explain all of the variation, as many components are needed as there are variables in the analysis. For example, for NAEP Reading at the 4th grade level, there were 1081 initial variables (Allen, Donoghue, and Schoeps, 2001). For the NAEP background questions for 4th grade Reading, 381 components (e.g., conditioning variables) explained about 90% of the variation in the data – 1081 components are needed to explain 100% of the variation. The reduction from 1081 coded variables to 381 conditioning variables implies that many fewer than 1081 variables could be used to predict the student proficiency and form the prior distributions if the variables were carefully selected. Using the conditioning variables typically reduces the number of variables used to predict the prior by a half to a quarter of the total number of variables.

A third complication is that the relationship between the conditioning variables and the proficiency estimates differs across states. Therefore, the conditioning process is done at the state level (or jurisdiction level) rather than for the national sample as a whole. For 1998 NAEP Reading, 44 separate conditioning analyses were run and the relationship between the conditioning variables and the proficiency estimates varied substantially from state to state. It is interesting that the number of conditioning variables used within a state was much less than the 381 identified from the entire sample. The number ranged from 278 (Florida) to 110 (Nebraska), suggesting that the number of background questions could be reduced substantially. The bottom line is that the procedures used in NAEP are complex and they involve a variety of trade-offs. The trade-offs related to the background questions will be considered in the next section.

Implications of the Estimation Procedure for the Selection of Background Variables

The work by Thomas (2002) indicates that there are two different cases that need to be considered when determining the effect of background questions on the estimation of NAEP proficiency distributions. These cases refer to the way test booklets are constructed. One case consists of booklets that share a common weighting of content. For NAEP tests that contain multiple content areas, this case has booklets that represent the content areas in the same way. The second case has booklets that vary substantially in content coverage. The second case usually occurs when blocks contain reading passages, writing prompts, or sets of items related to a common stimulus that do not allow common content distributions across booklets. For example, it may not be possible to administer all of the different types of writing prompts to a student in the time available.

Thomas (2002) showed that when the content is balanced across booklets, the background questions have less effect on the quality of estimates of the distributions than when the content is not balanced across booklets. This implies that the need for background questions is affected by the structure of the booklets. Booklets composed of independent items that require short units of time for response can more easily be made parallel in content then booklets composed of large, time consuming tasks. Thus, the cognitive test specifications and the background questions can not be considered independently. The specifications for both parts should be considered together.

The relationships among the background questions and the proficiency estimates also affect the quality of the estimates of the proficiency distributions. The stronger the relationship of the background variables to the proficiency estimates, the less error there is in predicting the prior performance for each examinee. All else being equal, better prior distributions will result in posterior distributions with smaller variation. The five plausible values sampled from the distribution will be more similar resulting in more precise estimation of the full distribution. This implies that the background variables should be selected to be related to the proficiency estimates.

The relationships among background variables are also important. To stabilize the results from the large number of coded background variables, composites of the variables are defined by a principal components analysis. However, the number of actual variables in the analysis is very large and the number of composite “conditioning variables” is also in the hundreds. This suggests that the process could be made more efficient by more carefully selecting background questions to increase the variance accounted for by each principal component and reduce the overall number. This will require a detailed analysis of the background questions that are related to each principal component to determine what is going into each composite. Then, the relationship between each principal component and the proficiency estimate should be studied to determine which composites are most highly related to the proficiencies. Such analyses should indicate the types of background questions that will be related to each other and the proficiencies. Those that are not contributing to the improvement in prediction can be eliminated from the analysis. It would seem that hundreds of variables are beyond what is needed for developing good prior distributions for the proficiencies.

However, there is also the possibility that eliminating some of the background questions could be counter-productive. Mislevy and colleagues (e.g., Mislevy, Beaton, Kaplan and Sheehan (1992)) have shown that if a variable is used in subsequent analyses, but not in the estimation of the prior distribution, estimates of statistics may be biased. The term "biased" as it is used here means that the estimates are systematically too high or too low. Therefore, some variables should be used for the estimation of the prior distributions even if they are not highly related to the proficiency measure.

Conclusions and Recommendations

The NAEP analysis procedures are extremely complex and involve hundreds of variables. This procedure is the closest thing to rocket science that exists in the field of educational assessment. Perhaps it is even beyond rocket science. Among all of the analysis procedures used in NAEP, probably the most complex part is using background variables to improve the estimates of student proficiency on the content measured by the NAEP cognitive tests. This analysis starts with the coding of the responses to categorical background questions such as the "race/ethnicity" question into a set of 0/1 variables that can be analyzed using statistical regression procedures. This is done because the "race/ethnicity" categories, and many other categorical variables, do not have any inherent numerical value or ordering. A person is either in a category or not in the category. To capture all of the possible responses that a student might give, the "race/ethnicity" item is coded into 36 0/1 variables for further analysis. This process greatly expands the number of background variables used in the analysis. Many of the background variables need to be coded in this way. Frequently, over 1,000 coded and original numeric variables are included in the analysis.

Any one of the set of background variables (both coded and those in the original form) provides relatively little information for predicting an examinee's proficiency. Some of these variables are fairly unreliable. Others are not strongly related to the proficiency being assessed. Also, some of the background variables provide redundant information and are highly intercorrelated with other background variables. Such variables cause problems in the statistical estimation procedures. To address these problems, composites of variables are formed using a statistical procedure called principal components analysis. These composites are called conditioning variables. Many fewer conditioning variables are used to predict the prior distribution than the total number of background variables.

There is a danger in dropping some background questions from the development of conditioning variables. If future analyses will be using those background questions and they have not been included in the development of conditioning variables, the statistical analyses may be biased, yielding estimates that are consistently higher or lower than the true values. This suggests that background questions can be identified for possible elimination on statistical grounds, but they should be reviewed to determine if they will be the focus of some subsequent analysis.

The research of Thomas (2002) into the relationship of the effect of the conditioning variables and the structure of the test booklets indicates that the conditioning variables are more important when each booklet does not represent the full range of content that is of interest. In such cases, assessments of specific kinds of content are missing for some students. The information from the conditioning variables seems to at least partially compensate for the missing data. This is the reason for using the complex estimation process in NAEP. This suggests that when the booklets for a content area balance the subcontent areas across all booklets, then the background variables are of reduced importance. In those cases, fewer background variables could be used. However, if it is not possible to balance the subcontent areas within all booklets, then more careful consideration should be taken before reducing the number of background questions.

The analysis of the use of background questions as part of the process for estimating the proficiency distributions led to the following recommendations. These recommendations largely stem from a goal to make the NAEP survey process more efficient by reducing the number of background questions that are administered. As suggested by Thomas (2002), time saved by reductions in background questions could be used to administer additional cognitive items to increase the accuracy of the direct assessment of what students know and can do.

Recommendation 1

Investigate the relationships between the current conditioning variables and the proficiency estimates. Currently, hundreds of conditioning variables are used to generate the prior distribution of performance for a student. It seems highly unlikely that all of those variables are making useful contributions to that process. The conditioning variables that are unrelated to the proficiency measures can be identified and the background questions that contribute to those composites can be identified. If the same background variables are identified across many of the state analyses, they are low impact and can be considered for exclusion. It would be prudent to do some research on the differences in the reported results when the low impact background questions are excluded and when they are not.

Recommendation 2

Identify background questions that are critical for reporting and secondary analyses. The work by Mislevy and others (e.g., Mislevy, Beaton, Kaplan and Sheehan (1992)) indicates that including background questions in the conditioning process reduces bias in results when those same background variables are used in subsequent analyses. Therefore, it is important to include background questions in the conditioning process that are certain to be used for further analyses.

Recommendation 3

Identify the conditioning variables that are most highly related to the proficiency measures and consider adding background questions that will enhance the characteristics of those conditioning variables. The estimation of proficiency is improved when conditioning variables are related to the proficiency measures. Through analysis of existing data, it should be possible to determine the types of background questions that will likely be related to the proficiency measured by the assessment. Then, it should be possible to identify new background questions that will be related to the proficiency measure. These questions can replace those that are poorly related to the proficiency measures.

Recommendation 4

Consider the content structure of NAEP booklets when evaluating the usefulness of background questions for proficiency estimation. If the booklets must be unbalanced in content because of the length of time needed to respond to test tasks or because of the stimulus related nature of the items, conditioning variables are more critical to the estimation of proficiency. Conditioning is less important if all booklets within a NAEP test have relatively parallel content. Therefore, the specifications for booklets and for background questions should be considered together as a single survey design process.

* * * * *

These recommendations are made with the following disclaimer. The recommendations are made based on information contained in the public literature about NAEP estimation processes. That literature is not sufficient to give a thorough understanding of the process. Only people who have done the work truly understand how this complex process works. It is possible that I have misunderstood some of the research and analysis results. If that is the case, I hope someone who is more knowledgeable about this process than I will point out any errors in my understanding of this process.

References

Allen, N.L., Donoghue, J.R., & Schoeps, T.L. (2001). *The NAEP 1998 Technical Report,* NCES 2001-509. Washington, DC: National Center for Education Statistics.

Beaton, A. & Zwick, R. (1992). Overview of the National Assessment of Educational Progress. *Journal of Educational Statistics, 17,* 95-109.*

Johnson, E. G. (1992). The design of the National Assessment of Educational Progress. *Journal of Educational Measurement, 29*(2), 95-110.

Johnson, E. G. & Rust, K. F. (1992). Population inferences and variance estimation in NAEP data. *Journal of Educational Statistics, 17,* 175-190.

Mislevy, R. J. (1991). Randomization-based inferences about latent variables from complex samples. *Psychometrika, 56,* 177-190.

Mislevy, R. J., Beaton, A. E., Kaplan, B. & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*(2), 133-161.

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics, 17,* 131-154.*

National Center for Education Statistics (1999). *The NAEP 1998 Reading Report Card: National & State Highlights.* Washington, DC: U.S. Department of Education.

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Thomas, N. (2002). The role of secondary covariates when estimating latent trait population distributions. *Psychometrika, 67*(1), 33-48.*

Thomas, N. & Gan, N. (1997). Generating multiple imputations for matrix sampling data analyzed with item response models. *Journal of Educational and Behavioral Statistics, 22*(4), 425-446.

Zeger, L. & Thomas, N. (1997). Efficient matrix sampling instruments for correlated latent traits: examples from the National Assessment of Educational Progress. *Journal of the American Statistical Association, 92,* 416-438.