STATISTICAL ISSUES IN DATA ANALYSIS FOR THE NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

Albert E. Beaton, Educational Testing Service

1. Introduction

The purpose of the National Assessment of Educational Progress (NAEP) is not to estimate the ability or achievement of individual students but to estimate the performance in particular subject areas of all children and young adults at specific age or grade levels and the performance of particular subgroups of students which are of interest. In the past, as well as at the present time, NAEP has aimed at providing information on a broad spectrum of appropriate and important skills and performance in the subject areas it has assessed. In fact, the Congressional Act authorizing NAEP specifies that:

Each learning area assessment shall have goal statements devised through a national consensus approach, providing for active participation of teachers, curriculum specialists, subject matter specialists, local school administrators, parents, and concerned members of the public.

As it should, the process by which NAEP produces its goal statements inevitably produces a large number of exercises to be assessed and constituencies to whom the particular results are to be reported.

In the 1983-84 assessment, the collection of objectives resulted in a number of exercises that was judged to be substantially more than what we could reasonably expect students to respond to. In fact, if a single student sat down to take all of the exercises and other questions that were administered to his or her age or grade level, the estimated time to complete the task would be over six hours. The inability to assign all of the exercises to every student results in collected data that cannot be organized into a single, complete rectangular data matrix that is easily analyzed using available statistical systems. The decision about which exercises are administered to which students is ultimately critical in circumscribing the analyses that can be performed after the data are collected. The NAEP staff designing this assessment considered having a broad coverage of educational objectives to be important, as in the past, and also considered the interrelationships among objectives, and even among subject areas, to be important. Since individual students are not being assessed, it is not an important issue whether or not each student receives a representative sample of the total exercise pool. These considerations, and others, led to a major change in the NAEP design.

The Education Commission of the States (ECS), which administered the NAEP grant since its inception in 1969 until this 1983-84 assessment, used multiple matrix sampling to assign exercises to students. Essentially, the assessment hours for an age group were divided into packages which would take a student about 3/4 of an hour to complete. Using this approach, the six hours of assessment exercises would result in eight packages. After the sample of students within a school was selected and brought to an assessment session, a particular package of exercises was distributed to all students within that session. A disadvantage of this procedure is that, although the interrelationships within a particular package can be studied, the interrelationships among packages cannot.

The Educational Testing Service (ETS), which is administering the grant this year, has chosen a complex variant of multiple matrix sampling called Balanced Incomplete Block (BIB) spiralling. This approach continues to allow the broad coverage of subject areas and also allows the study of the interrelationships among all exercises within and among subject areas. The basic idea is to divide up the total assessment time into small blocks. Each exercise block is then assigned to a number of assessment booklets such that each block of exercises is paired with each other block in some booklet. The booklets are then spiralled so that students in an assessment session are given different booklets. Using BIB spiralling, a large number of booklets must be created, but the interrelationships between objectives may be examined since each exercise is paired with each other exercise in some booklet.

The considerations in developing the design of the assessment instruments and the interplay between the amount of substantive coverage and the sample size will be discussed next. Then, the details of the BIB spiralling process as implemented for the NAEP, its perceived advantages and disadvantages, and the problems it creates for data analysis will be the subject of the rest of this paper.

2. Considerations in the NAEP Assessment Design

The design of any study is circumscribed by the amount of funds available, and thus the NAEP staff had to decide how to allocate its resources to allow as broad an assessment of its 1983-84 subject areas, reading and writing, as possible. The decisions that resulted in the final design were as follows:

1. Each student would be asked to participate for about 3/4 of an hour. To have a national assessment at all requires the cooperation of schools, and we felt, as did the ECS staff before us, that limiting the intrusion on individual students to about one class period would help us gain acceptance in the schools. The design originally called for 46 minutes of assessment for each student, but was extended to 48 minutes when a review of the early data showed that students were not reaching some important background and attitude questions.

2. The available funds were sufficient to gather data on about 30,000 students at each age/grade level. It should be noted that, under the terms of the grant, the Research Triangle Institute provided the sample of schools. Westat, who is the ETS subcontractor for sampling and field administration, reviewed the sample and studied some preliminary data collection plans to estimate the number of students who could be assessed for the available funds. 30,000 students at 48 minutes per student resulted in an expected total of 24,000 hours of testing time at each age/grade level.

3. Each exercise would be responded to by 2,600 students at each age/grade level. In past

assessments, around 2,500 to 2,600 students at each age level were targeted for each exercise. We felt that the efficiencies of BIB spiralling (see Hansen, et al., 1984) would allow us to reduce the number of students to about 2,000 without an increase in sampling error. However, we were committed to sample both the age levels which were sampled in the past (ages 9, 13, and 17) and also the grades into which most of those youths fell (grades 4, 8, and 11). We estimated that a sample of 2,600 at each age/grade level would result in a sample of about 2,000 at each age and also about 2,000 at each grade.

4. 5,000 students at each age/grade level would be set aside for assessment using a tape recorder. Data has been collected in national assessments since 1969, and we did not want to lose continuity with the already collected data. Since we were making a change from administration by tape recorder to pencil and paper administration, we felt that we needed to see what effect the method of administration had on the performance of students on assessment exercises. Therefore, a sample of 5,000 students was set aside for assessment using the same procedures as in the past.

5. There would be six minutes of questions common to all students. Some questions, such as racial/ethnic identification, are so important in the assessment that they must be asked of every student. At first, four minutes were allowed for such questions, but early experience required us to increase this section to six minutes.

6. Assessment exercises and other background and attitude questions would be grouped into blocks which would require 14 minutes to complete. These blocks would contain an average of 12 minutes of reading and writing exercises and two minutes of background and attitude questions. Thus, each student's 48 minutes would include the common questions (6 minutes) and three blocks of other assessment questions (14 minutes each). In terms of content, a student would spend 12 minutes on background and attitude questions and 36 minutes on reading or writing exercises.

7. Several long exercises, that could not be administered in 14 minute blocks, had to be included. Under the grant, the set of exercises to be used in this assessment was supplied by ECS and was somewhat larger than we could use. Our task, therefore, was to select exercises from this pool. Subject matter experts judged some of the longer exercises to be so important that they must be included, and they were accommodated by creating three double-length blocks (28 minutes).

It is immediately clear that a perfectly balanced incomplete block design is impossible, since the double-length blocks can not be paired within the 46 minute time limit. Although we could not assign two double-length blocks to any student, we could assign them in such a way that we could compare the double-length blocks indirectly through one or a chain of single-length blocks, and we did.

The final sample consisted of three parts, one of which received BIB spiralled booklets, a second received partially BIB spiralled booklets, and the third was a matrix sample which was assessed using tape recorders. The sample sizes and the amount of assessment time for the different samples are shown in Table 1.

The Balanced Incomplete Block (BIB) Sample

The booklets in the BIB design each contain the common block and three of the 19 single-length blocks assigned to this sample. The 19 blocks were assigned to booklets using a cyclic Youden rectangle (see Beall, 1971). This procedure required the formation and printing of 57 different booklets and assigned each individual block to precisely 9 different booklets. Each block is combined with each other block exactly once in this design, and thus each pair of exercises was assigned to some sample of youths. The block assignments were randomized. There are 57 block triplets, one for each booklet, which were a result of the randomization process. The assignment of blocks to booklets is shown in Table 2.

As shown in Table 1, this design called for each booklet to be administered to 288.9 different students and, since each block was in nine booklets, each block was therefore to be given to about 2,600 students, our target, at each age/grade combination. Altogether, this part of the design called for 288.9 students taking one of 57 booklets and thus 16,467 students in all. Looking at the age and grade samples separately, we expected each booklet to be administered to 222.2 youths at each age or grade level, thus each block to be administered to 2,000 youths resulting in a total age or grade sample of about 12,667.

The Partially Balanced Incomplete Block (PBIB) Sample

The booklets in the partially balanced design each contain the common block, a single-length block, and a double-length block. This design used seven blocks: three double-length blocks, two "new" single-length blocks, that are not used in the completely balanced design, and two "old" blocks that were also used in the other design. This design resulted in the formation and printing of six booklets. Two of the double-length blocks were combined with one of the new and one of the old blocks; the other double-length block was paired with both of the new blocks. The assignment of blocks to booklets is shown in Table 3.

The design called for each of these booklets to be administered to 1,300 youths and, since each of the new blocks were in exactly two booklets, each block was also administered to 2,600 youths. Altogether, the design called for 7,800 students taking a PBIB booklet. The design also met our objective of having about 2,000 students taking each exercise if we looked at the sample for a specific age, or 2,000 students if we looked at a particular grade.

The two booklets which contain a double-length block and one of the single-length blocks from the completely balanced sample result in an oversampling of these two single-length blocks since they are already adequately sampled in the BIB design. These two single-length blocks occur in 9 BIB booklets, each of which is administered to about 289 persons, and in one booklet of the partially balanced design, which is administered to 1,300 youths, and thus the targeted sample for each of these blocks was 3,900.

The Tape Sample

Four assessment booklets were designed for the tape sample and each was to be administered to a subsample of 1,250. Each booklet contained the

six minute common block and 42 minutes of cognitive exercises and background and attitude items. Since a tape recorder was used in administration, all students in an assessment session were assigned the same booklet.

Spiralling

The booklets from the completely balanced and partially balanced designs were then spiralled together for administration. To meet our targeted samples, we needed to administer nine BIB booklets for every two partially balanced booklets. Recall that there were 57 BIB booklets and 6 partially balanced booklets. This led to a cycle of two BIB iterations (114 booklets) and nine partially balanced cycles (54 booklets) for a total cycle of 168 booklets. Each type of booklet was randomized and then a string of 168 piles of about 300 booklets was formed in a chain of the form

1,2,58,3,4,59,5,6,60,7,8,61,9,10,62,11,12,63,
13,14,58,15,16,59,17,18,19,60,...,55,56,57,
59,1,2,61,3,4,62,...,55,56,57,63

where the numbers 1,2,...,57 represent booklets from the BIB design and 58,59,...,63 represent booklets from the partially balanced design.

The booklets were then placed in packages to be administered to different assessment sessions. The size of a package was important in making the design work, since we did not want any particular booklet to have a higher probability of being at the end of a package and thus be more likely not to be administered because of absenteeism or other reasons. It was decided to place 23 booklets in a package for each age/grade level. In this way, each booklet had the appropriate probability of being at any particular position within a package.

Implementation

At present, the final results of this design are not yet available, but the early returns seem to show that it was implemented effectively. A full report, showing actual sample sizes for each block, will be prepared when the information is available.

3. Advantages and Disadvantages

A large, complex assessment design such as used in the 1983-84 national assessment has a number of advantages and disadvantages, which should be mentioned.

Interrelationships among exercises

The purpose of the BIB spiral design was to allow the examination of the interrelationships of a large number of exercises, and it does. The final sample includes 19 14-minute blocks, 266 minutes in all, of exercises, and for any pair of exercises in these blocks there is a sample of youths who was presented both exercises. Thus, correlations can be computed among all the exercises in this part of the sample. The remaining 112 minutes of exercises are organized so that some, but not all, of the correlations can be calculated.

This design is in contrast to the multiple matrix design which was used previously. Given a fixed sample size, matrix sampling and BIB spiralling would administer any particular exercise to the same number of youths, but, by creating more booklets, the BIB design would pair the exercises in a block with many different blocks of exercises and thus increase the number of comparisons that could be made. The consequences of this are to have many correlations, but most based on a fairly small, although well selected, sample. In the design as implemented, correlations within a block are based on about 2,600 students (2,000 for an age or grade separately) but correlations between blocks are based on about 289 students (222 for an age or grade separately). Multiple matrix sampling allows the exercises within a larger booklet to be correlated, but does not allow any calculation of correlation coefficients among the exercises in different booklets.

The Cost of Complexity

Clearly, BIB spiralling is expensive in printing costs as well as in the costs of design talent and managing the details. Including the multiple matrix sampling that was done for this NAEP, there were 67 booklets created for each of the three age/grades being assessed, thus there were 201 booklets created in all. It is expensive to produce many booklets in small volumes. It was tedious to manage a task in which every detail had to be multiply checked. Another substantial cost was an intelligent data entry system since machine reading of the booklets was impossible.

The BIB spiralling had, however, reduced costs in some ways. The system was robust against failures in the field, since a serious biassing of results by having the exercise administrators use the wrong bundle of booklets was most unlikely and would affect the design very little. The loss of the tape recorder reduced costs in both preparation and administration of the assessment. Perhaps most importantly, as noted elsewhere, the BIB design also reduced the number of students needed to achieve a fixed standard error and thus allowed us to assess more exercises.

Tape Recorded Administration

Losing administration by tape recorder was not something that the NAEP staff wanted, but came about because of the BIB spiralling. It is clear that, when each student in an assessment session is taking a different booklet, a single tape recorder in the front of the room cannot read the questions aloud. We did not consider individual tape recorders.

The advantage of tape recorded administration is that it allows the separation of reading ability from the subject area being assessed. In a reading assessment, the instructions are tape recorded and the progress through the assessment is paced, although the reading exercises themselves are, of course, not read. In other subject areas, the exercises are read aloud so that students can respond to an exercise even though they may not be able to read it. This is clearly a desirable feature.

And yet, the utility of the NAEP is greatly enhanced by developing exercises that teachers or local or state personnel can readily administer to their students and then can compare the results to the NAEP sample. Teachers are not likely to simulate the tape recording, and thus any comparisons would be suspect. We know of no local or state assessments that have duplicated the tape recorder design. Thus, the tape recorder had the effect of setting the NAEP results apart from all other student assessments.

Sampling Efficiency

The efficiency of BIB spiralling is discussed elsewhere by Hansen et al. (1984), and will not be discussed in detail here. Basically, the advantage of BIB spiralling is that it presents a particular block of exercises to fewer persons in a school,but to more schools. In this way, the cluster effect is markedly reduced and thus the students are used more efficiently. Hansen et al. have estimated that, given reasonable assumptions, the required sample size to achieve a given standard error is reduced by about 20-25% by BIB spiralling, as compared to multiple matrix sampling; alternatively, the standard errors could be reduced by about 10-15% if the sample size were kept constant.

4. Statistical Issues

As mentioned above, BIB spiralling does not result in a complete, rectangular data matrix that can be analyzed using standard statistical systems nor does it generate data which are consistent with normal statistical methods. A few very basic problems are noted below, and final solutions to these problems have not yet evolved. We are, however, making some progress which may be of some interest.

Missing Data

Essentially, we would like to compute the correlations among the responses to all exercises, but we have not administered all exercises to all students. Since we could ask for only about 45 minutes of a student's time, some, in fact a substantial portion of the data we would like to have, is missing. The BIB spiralling design does allow separate estimates of all correlations among the exercises, but the estimates do not necessarily produce a matrix that is at least positive semi-definite. The correlations between two distinct exercise blocks are based on the responses to the booklet in which the two blocks appear; the correlations of either of these blocks with any other block will be based on a different booklet, unless the three blocks happen to be in the same booklet. Since each student receives only one booklet, the correlations between different pairs of blocks are usually based on different random samples of students. Each sample is subject to sampling error, and thus the correlations in a pair of blocks may be inconsistent with those in another pair. In processing some early data, we have found that the total correlation matrix is inconsistent with a complete data matrix.

A theoretically elegant way of producing a matrix that is at least positive semi-definite is by use of the EM algorithm (Dempster, Laird, and Rubin, 1977), but we fear that our data base is too large to make this approach practical for the general problem, although we may use it for subsets. 57 patterns of missing data with around 500 variables is just too big for even our virtual memory, and the slowness of the EM algorithm makes it possibly unaffordable. We are also exploring a method proposed by Wingersky (1984) which is a variation of MINRES algorithm. We are hopeful that this will result in a reasonable and practical solution to our inconsistent matrix problem.

Dimensionality

We would like to summarize the reading exercises by one, or a few, composites of the collected data -- to construct a composite of reading achievement which can be measured from assessment to assessment and the results compared for trends. We have proposed using modern item response theory models to do this, and have at this time been reasonably successful. Item response theory assumes that the observed data are generated from an underlying one dimensional model, and there is no widely accepted test for unidimensionality. This issue is especially important in NAEP since some assert that achievement data can logically be presumed to be multi-dimensional.

NAEP is taking a multi-pronged approach to the dimensionality issue including supporting some new methods which have just been proposed. We are using the classical approach, which is to compute tetrachoric correlations among the exercise responses, and to examine the factorial structure of this matrix. We are also using several other newly proposed methods which seem promising and on which we will report later.

Parameter Sampling Error Estimation

Given the sampling weights, the estimation of population proportions and averages is relatively straightforward, but the estimation of variances and covariances is less so, and the estimation of sampling errors is quite complex. Some early simulations led us to the jackknife as a general procedure for estimating error variances, and we have programmed a general procedure for use with many NAEP statistics. The results so far seem good -- the sampling error is smaller than we feared -- but we do not yet feel comfortable enough with the accuracy of our results to report on them. The results will be reported at a later time.

References

Beall, Geoffrey: Change-Over Experiments in Practise, 1971, Princeton, ETS RB71-38

Dempster, A. P., Laird, N. M., and Rubin, D. B.: Maximum Likelihood Form Incomplete Data Via the EM Algorithm. J. of the Roy. Stat. Soc., Ser. B, 1977, 39, 1-38

Hansen, M. H., Tepping, B. S., Lago, J. A., and Burke, J.: National Assessment of Educational Progress (NAEP) -- The Sample and Data Collection Design for Year 15. Paper presented at the 1984 meeting of the A.S.A.

Wingersky, B.: Innovation in Multivariate Statistics. Paper in preparation.

Robert F. Boruch of Northwestern University was a discussant of this paper at the 1984 ASA Conference.

TABLE 1
SAMPLE DESIGN SUMMARY

Sample for Age/Grade	Blocks: Single	Blocks: Double	Booklets	Youth per Booklet	Youth per Block	Youth per Sample	Assessment Time in Minutes: Common	Subject Matter	Other	Total
BIB -										
Age & Grade	19	0	57	156	1,400	8,867	6	228	38	272
Age Only				67	600	3,800				
Grade Only				67	600	3,800				
Total	19	0	57	290	2,600	16,467	6	228	38	272
PBIB -										
Age & Grade	4*	3	6	700	1,400	4,200	6	120	20	146*
Age Only				300	600	1,800				
Grade Only				300	600	1,800				
Total	4*	3	6	1,300	2,600	7,800	6	120*	20	146*
Tape-(Age Only)	12	0	4	1,250	1,250	5,000	6	144	24	174
TOTAL FOR AGE/GRADE	21	3	67	-	-	29,267	6	324	54	384
TOTAL OVER ALL AGE/GRADE	63	9	201			87,801		**		

* 2 Single blocks are duplicated in BIB Sample

** Total assessment time depends on common blocks across age/grade

TABLE 2
NAEP BOOKLET DESIGN
BIB SPIRAL SAMPLE
(19 X 3 X 57 Cylic Youden Rectangle)

Booklet	ORIGINAL DESIGN Part 1	Part 2	Part 3	PERMUTED DESIGN Booklet	Part 1	Part 2	Part 3
1	1	2	7	1	19	7	11
2	2	3	8	2	1	11	15
3	3	4	9	3	4	1	19
4	4	5	10	4	3	18	5
5	5	6	11	5	3	1	8
6	6	7	12	6	7	6	8
7	7	8	13	7	10	17	13
8	8	9	14	8	17	12	6
9	9	10	15	9	14	13	11
10	10	11	16	10	6	4	2
11	11	12	17	11	5	12	1
12	12	13	18	12	18	8	2
13	13	14	19	13	12	10	4
14	14	15	1	14	19	13	9
15	15	16	2	15	12	19	3
16	16	17	3	16	3	11	16
17	17	18	4	17	8	5	17
18	18	19	5	18	3	15	6
19	19	1	6	19	11	18	10
20	1	3	11	20	13	2	5
21	2	4	12	21	13	3	4
22	3	5	13	22	16	10	8
23	4	6	14	23	11	8	4
24	5	7	15	24	1	18	17
25	6	8	16	25	11	9	17
26	7	9	17	26	19	6	16
27	8	10	18	27	3	10	9
28	9	11	19	28	14	9	18
29	10	12	1	29	16	14	4
30	11	13	2	30	2	16	9
31	12	14	3	31	14	19	8
32	13	15	4	32	2	12	11
33	14	16	5	33	7	17	14
34	15	17	6	34	7	14	5
35	16	18	7	35	18	16	12
36	17	19	8	36	2	1	14
37	18	1	9	37	10	7	1
38	19	2	10	38	14	6	10
39	1	4	8	39	15	18	19
40	2	5	9	40	5	6	11
41	3	6	10	41	8	12	9
42	4	7	11	42	9	5	4
43	5	8	12	43	6	9	1
44	6	9	13	44	2	7	3
45	7	10	14	45	15	2	10
46	8	11	15	46	18	6	13
47	9	12	16	47	15	16	5
48	10	13	17	48	2	17	19
49	11	14	18	49	15	12	14
50	12	15	19	50	17	15	4
51	13	16	1	51	7	17	16
52	14	17	2	52	18	7	3
53	15	18	3	53	8	15	13
54	16	19	4	54	19	5	10
55	17	1	5	55	12	7	13
56	18	2	6	56	1	13	16
57	19	3	7	57	7	9	15

TABLE 3

NAEP BOOKLET DESIGN

PBIB SAMPLE

Booklet	Block 1	Block 2
58	B_{18}	D_1
59	B_{20}	D_1
60	B_{19}	D_2
61	B_{21}	D_2
62	G_{20}	D_3
63	B_{21}	D_3

where

B_{18} and B_{19} are the single blocks which are also in the BIB spiral sample

B_{20} and B_{21} are the single blocks which are in the PBIB sample only

and

D_1, D_2, and D_3 are the double length blocks

TABLE 4

ESTIMATED SAMPLE SIZES

FOR ALL PAIRS OF BLOCKS

FOR EACH AGE OR GRADE

	BIB Sample								Package Sample				
	B_0	B_1	B_2	B_3	...	B_{17}	B_{18}	B_{19}	B_{20}	B_{21}	D_1	D_2	D_3
B_0	18,667	2,000	2,000	2,000	...	2,000	3,000	3,000	2,000	2,000	2,000	2,000	2,000
B_1		2,000	222	222	...	222	222	222	0	0	0	0	0
B_2			2,000	222	...	222	222	222	0	0	0	0	0
B_3				2,000	...	222	222	222	0	0	0	0	0
⋮						⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
B_{17}						2,000	222	222	0	0	0	0	0
B_{18}							3,000	222	0	0	1,000	0	0
B_{19}								3,000	0	0	0	1,000	0
B_{20}									2,000	1,000	1,000	0	1,000
B_{21}		SYMMETRIC								2,000	0	1,000	1,000
D_1											2,000	0	0
D_2												2,000	0
D_3													2,000

B_0 is the common block

B_1 to B_{19} are the BIB Spiral blocks

B_{18} to B_{19} are the two single blocks which are in both BIB and Package Samples

B_{20} and B_{21} are the two single blocks which are in the Package Sample only

D_1 to D_3 are the double blocks

Note: These estimated sample sizes are appropriate for either the grade or age samples. To estimate the total number of students, including both age and grade samples, multiply these numbers by 1.3.