DESIGN EFFECTS IN THE NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS SURVEYS

J. R. Chromy, R. P. Moore, and Anne Clemmer
Research Triangle Institute

I. INTRODUCTION

The National Assessment of Educational Progress [1] is a long-range study of the knowledge, skills, understandings, and attitudes of certain specified subpopulations of young Americans. National Assessment's two major goals are (1) to make available the first census-like data on performance of these specified subpopulations on exercises designed to measure specific educational objectives within selected subject matter areas, and (2) to measure changes in performance on these exercises over time.

To achieve these objectives, a system of annual surveys of four national age groups has been developed. In each of the first three annual surveys, special exercises have been developed in two or three subject matter areas. The Year 01 survey, begun in 1969, involved Science, Citizenship, and Writing; the Year 02 survey involved Reading and Literature. The first opportunity to measure change over a period of time will occur in Year 04 with the reassessment of Science. Some design effects [2] for Science and Writing exercises from Year 01 are presented in this paper. In addition, some preliminary results from selected Year 02 exercises are presented. The design effects in this paper relate to national age group estimates or smaller subpopulation estimates of the proportion of the subpopulation answering specific exercises correctly.

The emphasis in National Assessment reporting on the performance of subpopulations on individual exercises is unique in educational measurement. Because of this feature, it was not necessary to require each respondent to participate in all exercises in any annual survey, but it was necessary only to require that a probability sample of the target population be obtained for each exercise. As a result, the exercises for each year of National Assessment were assembled into several packages and each respondent participated by completing only one of the packages on a probability basis. For respondents assessed through the school sampling frame, packages were administered on a group basis and on an individual interview basis. Individual administration methods were used with respondents contacted through the household sampling frame.

Because of the multiple package (questionnaire) approach to National Assessment and other special requirements for adequately representing subpopulations defined by region, by community characteristics, and by socioeconomic status characteristics, the sample design for National Assessment was extremely complex when considered in its entirety and can only be accurately described by going into considerable detail. A brief description of the Year 01 sample design and estimation procedures, however, is considered necessary before the Year 01 design effects may be discussed.

II. THE YEAR 01 SAMPLE DESIGN

The four target age populations were 9-year-olds, 13-year-olds, 17-year-olds, and young adults of ages 26-35 residing in the United States. The 9-, 13-, and 17-year-olds enrolled in elementary and secondary schools were sampled through a school sampling frame. Out-of-school 17-year-olds and young adults of ages 26-35 were sampled from a household sampling frame; out-of-school 17-year-old data are not included in this paper. In general, the sample design for either frame could be described as a multistage stratified design.

A sample of 208 primary sampling units was selected. Primary sampling units were defined conceptually within larger listing units consisting of cities, counties, portions of counties, or groups of counties. In Year 01, the same sample draw of primary sampling units was used for both the school and household frame. The subsequent stages of the sample were selected independently within these primary sampling units. The primary sample was stratified by region, by community-size characteristics, and by income characteristics. In most cases, two primary sampling units were selected per stratum; in a few smaller strata, only one primary sampling unit was selected.

The sample design for a particular National Assessment package can be described more easily than the entire design for all packages. For the school sample, the design for a group-administered package involved one group of twelve students randomly selected from one school in each primary sampling unit. The design for an individually administered package involved nine students per primary sampling unit (PSU) selected from (typically) 2 to 5 schools.

In the household sample, the second-stage sampling units were area clusters of housing units. Ten second-stage sampling units were selected in each PSU. The package sample size per second-stage sampling unit under this design was a random variable, and averaged less than one. Table 1 gives a quick view of the sample design for both sampling frames.

An attempt was made to sample from certain subpopulations at a higher rate than from others in order to provide adequate subpopulation sample sizes for reporting purposes. The disproportionate sampling methods were applied at both the first stage and subsequent stages of the sample selection whenever the necessary identifying data were obtainable.

As previously stated, the estimates reported from the National Assessment data are the

Table 1

Year 01 Sample Design Summary

Sampling frame	Number of PSU's	Within-PSU Sample
School Frame		
Group-administered packages	208	12 students selected from one school
Individually administered packages	208	9 students selected from 2 or more schools
Household Frame		
Individually administered packages only	208	10 second-stage area clusters. Respondent sample size per package is a random variable.

estimated proportions of subpopulations who would respond correctly to specific exercises. These estimates (P-values) were computed as combined ratio estimates. The denominator of the ratio estimate was an estimate of the size of the subpopulation of interest, and the numerator was an estimate of the number of persons in the subpopulations who would perform in an acceptable manner on the given exercise.

The Horvitz-Thompson estimator was used for estimating both the numerator and denominator values in the ratio estimate; each sample response was given a weight or expansion factor equal to the inverse of the respondent's probability of selection for the package containing the particular exercise. These weights were further adjusted, where necessary, to account for nonresponse.

Variances of these estimates of proportions were estimated by using a first-stage jackknife estimator and ignoring the finite correction factor. Since the finite correction factors were small, this estimator would be expected, on the average, to give a small overestimate. Some collapsing of strata was also necessary to include the variance contributions from strata where only one primary sampling unit had been selected.

III. YEAR 01 DESIGN EFFECTS

National design effects were estimated for 149 science and writing P-values. The median design effect estimate for the 149 exercises examined was 2.38, with the majority of the 149 design effects falling between 1.50 and 3.00. As Table 2 shows, 82 percent of the design effects were 3.00 or less, 87 percent were 3.50 or less, and 94 percent were 4.00 or less. Table 3 shows median design effects and ranges in design effects for various subgroups of national design effects classified by age group, administration mode, and subject matter area.

The design effects for group-administered exercises were higher than those for individually administered exercises due to more clustering of the sample respondents. As stated, each group package was administered once in each PSU to a group of 12 students selected from a single school. For each individual package, 9 respondents per PSU were selected from several schools.

The design effects for 13-year-olds were smaller than those for 9-year-olds, while the 17-year-old exercise design effects were smaller than those for either 9- or 13-year-olds. This should be expected since the older the students are, the more likely they are to be enrolled in larger schools serving larger, more heterogeneous populations. It may be that high schools are more heterogeneous in terms of students than are junior high schools, and both junior high and senior high schools are more heterogeneous than the elementary schools.

The design effects for adults 26 to 35 years of age were about equal to those for 9-year-olds, possibly reflecting a similar intracluster correlation for the household sample due to small, compact clusters and variable housing patterns within PSU's.

There was no apparent difference in design effects for science exercises as compared with writing exercises. The comparison is difficult because of the small number of group-administered writing exercises and the fact that no individually administered science exercises were examined in this study.

Table 2

Distribution of National Design Effects

Design Effect	Number	Percent
< 1.00	1	1%
1.00 - 1.50	16	11%
1.51 - 2.00	29	19%
2.01 - 2.50	43	30%
2.51 - 3.00	32	21%
3.01 - 3.50	8	5%
3.51 - 4.00	10	7%
4.01 - 4.50	5	3%
4.51 - 5.00	3	2%
> 5.00	2	1%
Total	149	100%

Table 3

Median Design Effects for National P-Value Estimates

Age	Administration Mode	Subject Area	Number of Exercises	Median Design Effects	Range of Design Effects	Mean Number of Respondents
9	Group	Science	30	2.68	1.92-4.94	2,442
13	Group	Science	27	2.26	1.31-6.01	2,415
17	Group	Science	10	1.81	.90-2.51	2,122
17	Individual	Science	1	1.13	1.13	579
26 to 35	Individual	Science	16	2.57	1.38-4.08	878
9	Group	Writing	24	2.81	1.51-3.80	2,426
13	Group	Writing	5	4.36	1.93-10.88	2,416
9	Individual	Writing	13	2.21	1.45-2.68	1,817
13	Individual	Writing	23	1.89	1.24-2.88	1,863

Tables 4, 5, and 6 show median design effects for subpopulations defined by regional strata, and for sex and size of community subpopulations defined by poststratification. The median design effects for subpopulation estimates are of about the same magnitude or slightly smaller than the median design effects for national estimates.

The median design effects for 9- and 13-year-old writing exercises tended to be highest for the Southeast region (Table 4).

No consistent differences were noted in the median design effects for males and females (see Table 5).

Table 6 shows median design effects for size of community (SOC) subpopulations defined by poststratification. As with national design effects, the median design effects for SOC subpopulations are higher for group-administered exercises than for individually administered exercises. There is possibly a tendency for the metropolitan and urban area median design effects to be smaller than those for more sparsely populated medium city and rural (small place) subpopulations.

The design effects shown in this paper reflect the combined effects of clustering of the sample, unequal weighting of sample respondents, stratification, and other sample design and estimation factors. The effect of unequal weighting of sample respondents has been estimated to be from 1.3 to 1.6, depending upon the exercise.

IV. SOME PRELIMINARY YEAR 02 RESULTS

Some major revisions in the National Assessment sample design occurred in Year 02. The first principal change involved doubling the within-primary sampling unit sample and halving the number of primary sampling units. The planned number of administrations per individual package was increased to 20 per primary sampling unit.

The second change involved the use of controlled selection of the primary sample to permit stratification by State as well as by the previously discussed set of stratification variables. Due to this change, some approximate variance estimation procedures were required and only partial results have been obtained.

An approximate variance estimation formula based on the Yate-Grundy variance estimator was used to study design effects in one region. To isolate the design effect associated with clustering from the design effects associated with unequal weighting, stratification, and other design factors, the same variance estimation formula was also applied to systematic half-samples of the ordered data file for selected items. The variance estimates from the two half-samples were averaged. Design effects were then computed for both the whole sample and the systematic half-samples.

If a positive intraclass correlation existed, the design effect would be expected to be larger using the whole sample data. With a negative intraclass correlation, the opposite result would hold. Table 7 gives some results for ten selected exercises. A measure of homogeneity is also given for the group-administered exercises. This measure, based on the ratio of the change in design effect to the change in cluster size, should behave similarly to the intraclass correlation coefficient under a self-weighting cluster design. If this condition continues to hold under more complex sample designs, then the method described provides a convenient means of describing the behavior of the variance as cluster size is varied.

Table 4

Median Design Effects for Regional Subpopulation P-Value Estimates

(9- and 13-Year-Olds Only)

Age	Administration Mode	Subject Area	Number of Exercises	Median Design Effects by Region			
				Northeast	Southeast	Central	West
9	Group	Writing	24	1.89	2.93	2.32	2.65
13	Group	Writing	5	3.05	3.65	3.50	2.65
9	Individual	Writing	13	2.34	1.30	1.85	2.17
13	Individual	Writing	23	1.64	2.11	1.61	1.35

Table 5

Median Design Effects for Sex Subpopulation P-Value Estimates

Age	Administration Mode	Subject Area	Number of Exercises	Median Design Effects by Sex	
				Males	Females
13	Group	Science	20	2.57	2.25
26 to 35	Individual	Science	15	2.08	2.20
9	Group	Writing	24	2.74	2.54
13	Group	Writing	5	2.95	4.38
9	Individual	Writing	13	2.27	2.03
13	Individual	Writing	23	1.80	1.84

Table 6

Median Design Effects for Size of Community (SOC)
Subpopulation P-Value Estimates

Age	Administration Mode	Subject Area	Number of Exercises	Median Design Effects for SOC Categories			
				Big City	Urban Fringe	Medium City	Small Place
9	Group	Science	30	2.26	2.01	2.56	3.38
13	Group	Science	27	2.43	2.20	2.14	1.90
26 to 35	Individual	Science	16	1.91	2.25	1.47	1.86
9	Group	Writing	24	2.04	2.18	2.41	2.86
13	Group	Writing	5	3.79	2.95	3.82	3.69
9	Individual	Writing	13	1.75	1.97	2.66	1.91
13	Individual	Writing	23	1.22	1.37	1.75	2.38

Table 7

Year 02 In-School Design Effects For 10 Selected Exercises

Administration Mode	Age	Exercise	Sample Size	Estimated Design Effects		An Approximate Measure of Homogeneity
				Half-Sample	Whole Sample	
				(D_1)	(D_2)	(Δ)*
Group	9	1	618	2.90	5.81	.485
		2	618	.82	.61	-.035
	13	1	655	2.57	3.78	.202
		2	655	1.89	2.57	.113
	17	1	603	2.98	4.52	.256
		2	603	.15	.69	.090
Individual	9	1	549	1.07	.74	---
		2	549	1.42	2.36	---
	17	1	541	1.79	1.34	---
		2	535	.30	2.90	---

*$\Delta = (D_2 - D_1)/(12 - 6)$ for group-administered exercises only.

The only immediate conclusion one can draw is that the estimates of standard errors and design effects vary considerably from exercise to exercise. Further study is needed to determine if the wide variation in design effects is a real phenomenon or whether such results are obtained empirically only because the variance estimates themselves have a high variability.

REFERENCES

[1] A project of the Education Commission of the States.

[2] Leslie Kish, Survey Sampling. New York: John Wiley and Sons, 1965.