LONGITUDINAL COMPARISONS AMONG SURVEY STATISTICS WITH SPECIAL REFERENCE TO THE NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

Edward C. Bryant and Morris H. Hansen, Westat, Inc.

1. Introduction

The purpose of this paper is to discuss comparison of survey statistics across time, where the objective is to measure changes in knowledge, skills, and attitudes. While the problem may be encountered in a variety of settings, we are concerned primarily with measuring "progress" in the context of the title of the project, "National Assessment of Educational Progress" (NAEP).

That project administers "exercises" to a sample of 9-, 13-, and 17-year-old children and to young adults (26 to 35 years of age). The exercises are usually scored as correct or incorrect, but some are scored into multiple categories. The aggregative measure for an individual exercise is a "p-value," which represents the percentage of persons whose answers are correct, or whose answers fall into (let's say) a given attitudinal response category. The p-values are not aggregated into test scores for an individual person although, as we shall see, some aggregation across exercises for groups of people is implied in summary measures used for generalizations about progress (or lack of it) over time.

After the first administration, some of the exercises were "released," i.e., made public along with their p-values. The others were not disclosed and were retained for use in future administrations to measure change in knowledge, skill, or attitude (by comparing p-values). There is concern that released exercises might be picked up and "taught to" or used in teaching, and consequently not be acceptable for measuring change. In the following sections we discuss procedures and problems in obtaining comparability of exercises and scoring and in increasing the sensitivity of comparisons across time by adjustment for background factors.

2. Comparability of Exercises

For approximately the half of the exercises given during the first administration that were not released, the difference between the p-values at the second administration and the p-values at the first administration is a measure of educational progress. Such measures are, of course, subject to sampling error. They may also be subject to bias:

a. if the scoring standards change,

b. if changes in method of administration affect the p-values,

c. if there is a failure in security of unreleased exercises, and

d. if the exercises themselves are related to temporal issues.

We will discuss changes in scoring standards in the next section. An important objective of administration procedures is to minimize the effect of the method of administration on exercise outcome. Changes in procedures, either to improve them or to cut costs, will have the inevitable effect of reducing comparability. Whether it is feasible to introduce changes by splitting the sample into halves, one half receiving the old procedure and the other the new, depends upon cost and the availability of this feature in the design.

The security of unreleased exercises probably is not of major concern as long as NAEP can retain its isolation from political and funding issues. NAEP has sometimes been criticized because the earth doesn't tremble when its outcomes are reported. This probably fosters comparability across time.

As an example of the relationship of an exercise to temporal things, consider the question: "What is the name of the President of the United States?" Here, the question may remain the same, but the p-value of this exercise may be expected to change through the office-holding period and may be substantially different for a President who is much in the news than for a President who is less visible.

We turn now to the question of release of exercises after the second round. Some alternatives are:

> Alternative 1. Retain the unreleased exercises from the first round indefinitely, using them at each round to measure progress.

This procedure would seem to produce the maximum comparability, but has some obvious shortcomings. First, it puts additional stress on the matter of security and, in case of any doubts about security, confounds the measures of progress with possible breaches in security. Second, there will be substantial interest in identifying the exercises that show extreme gains or extreme losses in p-values, but this cannot be done under this plan. One can, of course, identify the objectives with which the extreme gains or losses are associated, but this disclosure may not be satisfactory to the educational community. We believe that, as a practical matter, it would not be feasible to apply this procedure -- an effort to take this course is likely to prove exceedingly difficult if not impossible to follow.

> Alternative 2. Release, at the second round, all exercises that were unreleased after the first round, retaining as unreleased the new exercises at the second round.

This plan retains all exercises for two rounds with half of them "expiring" and therefore being released at each round. Thus, a link with the previous round is always available with half of

the exercises, and comparison with earlier rounds can be made by chaining the links. Implied in this approach is the ability to aggregate exercises or to pair them in some meaningful way.

Suppose one pairs an unreleased exercise at time t with a released exercise, where the pairing is made within common objectives and similar difficulty. We will refer to the set of possible exercises that are eligible to be paired as an exercise family. Then it is possible to measure "progress" with respect to a common objective for an exercise and all of its "ancestors" as follows.

Consider an exercise e that is administered at time $t-1$ and time t, and released after it is administered at time t, and let $p_{(t-1)ue}$ and p_{tre} be the p-values for the same exercise on the two dates, the subscript u denoting "unreleased" and r denoting "released." For convenience, we will let e represent an exercise used at times t and $t-1$ and all of its predecessors in the chain, which are presumed to be samples of the same exercise family. Then, a comparison of

$$\tilde{p}_{te} = p_{ore} \frac{p_{1re}}{p_{0ue}} \cdots \frac{p_{(t-1)re}}{p_{(t-2)ue}} \cdot \frac{p_{tre}}{p_{(t-1)ue}} \qquad (1)$$

with p_{ore} indicating gain (or loss) over time, subject, or course, to a sampling error. However, if time-linked p-values in a given subject matter show an upward movement in their distribution, measured perhaps by the median of the linked p-values, one might have some confidence that gain or progress was being demonstrated. If we assume that the exercises used in the chain are independently sampled from an exercise family, and that the persons to whom the exercises are administered are independently selected samples, the relative sampling error of r, the ratio of the time-linked p-value of released exercises at time t ($\tilde{p}_{tre}$ to p_{ore}) is approximately

$$\sigma_r^2 = 2tR^2V^2(1 - \rho)$$

where

R = the expected value of r;

t = the number of chained time comparisons, that is, the number of ratios of the type $r_t = p_{tre}/p_{(t-1)ue}$ in estimate (1);

$V^2 = V_b^2 + \frac{V_w^2}{t}$ is here assumed to be approximately constant over time;

V_b^2 = the relvariance between the expected p-values for the exercises in the exercise family;

V_w^2 = the average relvariance among students within exercises for the exercise family; and

$\rho = (V_{b12})/V^2$ with V_{b12} equal to the relcovariance of expected p-values for an exercise in two adjacent years.

The definition of these terms might be made clearer by the following illustration for two adjacent years, and with the simplifying assumption that the students taking the exercise are a simple random sample of eligible students. The observed p-value for exercise e at time t is p_{te} and its expected value is P_{te}.

Exercise (e) Within the Family	Expected p-value for Individual Exercises in the Family in Year	
	1	2
1	P_{11}	P_{21}
2	P_{12}	P_{22}
3	P_{13}	P_{23}
4	.	.
.	.	.
.	.	.
M	P_{1M}	P_{2M}

(We illustrate M as finite, but it may be regarded as indefinitely large.)

The average relvariances and relcovariances above are further defined as follows:

Average:

$$P_{t.} = \sum_{e}^{M} P_{te}/M$$

Relvariance between expected exercise p-values (for simplicity, assumed to be approximately equal over time):*

$$V_b^2 = \sum_{e}^{M} (P_{te} - P_{t.})^2/MP_{t.}^2$$

Relvariance within exercises (for simplicity assumed to be approximately equal over time):*

$$V_w^2 = \sum_{e}^{M} P_{te}Q_{te}/MP_{t.}^2$$

Relcovariance of expected p-values:

$$V_{b12} = \frac{\sum_{e}^{M} (P_{(t-1)e} - P_{(t-1).})(P_{te} - P_{t.})}{MP_{(t-1).}P_{t.}}$$

* Actually, if $P_{2.}$ is substantially greater or less than $P_{1.}$, it may be unreasonable to assume that the V_b^2 and V_w^2 are equal, but the amount of such variation in V_w^2 is not important for our present purposes.

Estimation of variances would call for use of two or more exercises from an exercise family at each time.

Alternative 3. Create a large pool of exercises, stratified by objective and (possibly) difficulty, and develop a rotation plan for adding (and releasing) a scheduled proportion of exercises at each round.

This approach differs from Alternative 2, above, among other reasons, because it involves creating a substantial pool of exercises and doing stratified random selection from the pool. In Alternative 2, we simply assumed that presumably comparable exercises are identified and chained in subsequent tests, without necessarily creating a family of exercises in advance and using an explicit random procedure for selection of exercises. We believe Alternative 3 (or a modification of procedures along these lines) has important advantages.

There is clearly an assumption of aggregation of exercises in this approach. That is, one assumes that there is a parameter representing change in knowledge, skills, or attitudes in a given subject matter, which can be estimated by average exercise scores. The estimate would have two components. The first component would be comprised of the difference in estimated average p-values of identical exercises given at two dates, presumably using weighted averages. The second component would consist of differences in average p-values of exercises that were different at the two dates, but that were drawn from the same exercise pool. Thus, an overall estimate of gain (or loss) could be expressed as

$$\bar{d} = z\bar{d}_c + (1 - z)\bar{d}_u \qquad (2)$$

where z is chosen so as to minimize the variance of $\bar{d}$, $(0 \leq z \leq 1)$,

$$\bar{d}_c = \sum_j w_j(P_{ctj} - P_{c(t-1)j}) , \qquad (3)$$

P_{ctj} denotes the p-value for the j^{th} common exercise at time t, the w_j are assigned weights, and

$$\bar{d}_u = \frac{\sum_j w'_j P_{utj}}{\sum_j w'_j} - \frac{\sum_j w''_j P_{u(t-1)j}}{\sum_j w''_j} \qquad (4)$$

where P_{utj} is the p-value of the j^{th} uncommon exercises at time t.

The variance and covariance estimates could be developed along the lines discussed above.

3. Comparability of Scoring

As we have pointed out before [1, 3], it is not a simple matter to maintain comparability of scoring across time for subjective exercises such as writing exercises or performance exercises -- playing of musical instruments, singing, etc. One way to obtain comparability is to have exercises given earlier rescored by persons who are scoring current performance. Then, assuming proper control can be maintained on other factors, one can obtain a set of p-values on exercises common to both administrations that are comparable. It may be necessary to photocopy writing exercises, for example, to get materials of equal readability.

4. Adjustments to Increase Comparability Across Time

Inherent in the adjustment process is the effort to define reasonably homogeneous subgroups of the population whose performance or outcome at a subsequent time can be compared with that of a similar group at an earlier time, such that if changes in average performance have occurred, it is reasonable to infer that the change reflects some real changes in performance and not simply changes in the composition of the subgroup. Such a subgroup might be "13-year-old Southern rural black males, neither parent completed high school." We assume that the characteristics that define the group (age, sex, geographic region, urbanization, education of parents, and race) have relatively stable definitions over time and that the classification of an individual as either in the group or not in the group is substantially error-free. Of the characteristics listed in the example, only degree of urbanization and education of parents are subject to change over time, and that change is likely to be slow enough for a five-year period that one need not be greatly concerned about such change on the comparability of classification.

The fact that one can identify similar groups over time is quite important. There are other variables, however, such as community and school variables, occupation of parents, items in the home, and other indicators of socioeconomic status (SES) that have been found to be useful in "adjusting" educational outcomes in order to make comparisons among population subgroups at a given date [3]. These variables typically do not remain stable over time -- a $10,000 income in 1974 does not represent the same thing as a $10,000 income in 1970, owning a color TV set does not have the same meaning in 1974 as it had in 1970, and so on. The question we address here is whether such background factors can be used successfully in increasing comparability of outcome measures across time.

The comparison problem can be explained with reference to the cells of the following table:

	Group 1	Group 2	Both Groups
Time 1	P_a	P_b	P_{a+b}
Time 2	P_c	P_d	P_{c+d}

in which the appropriate p-value is shown in the table.

During any one administration of the tests, one is interested in comparing one group with another, let's say, in comparing P_c with P_d.

The fact that NAEP usually compares one group against total U.S. performance, which includes the group of interest (i.e., P_c with P_{c+d}), tends to reduce the magnitude of the differences, but does not change the fundamental nature of the comparison. In NAEP, one is also interested in comparing performance of a defined group at time t with performance at time t-1, e.g., P_c with P_a, or P_d with P_b. These constitute measures of group gain and, of course, one is also interested in measuring overall gain, e.g., comparing P_{c+d} with P_{a+b}. Finally, one is interested in comparing gains by groups, such as $P_c - P_a$ with $P_d - P_b$.

Most of the literature has concerned itself with reducing bias in comparisons at one point in time, such as P_a with P_b or P_c with P_d. Adding the time dimension introduces some complexities, as we shall see.

We consider first the comparison of group means without adjustments for other variables. Let $\bar{y}_{gt}$ denote the mean outcome score (in National Assessment terms, a p-value) for group g at time t. Then, assuming that problems in comparability of exercises and in comparability of scoring have been solved (see above), the difference between the group means

$$d_g = \bar{y}_{gt} - \bar{y}_{g(t-1)} \tag{5}$$

is a meaningful measure of the gain in achievement for group g if the group is reasonably comparable at both dates, as discussed earlier.

An unadjusted aggregate measure of change for all groups combined is

$$\bar{d} = \bar{y}_t - \bar{y}_{(t-1)} \tag{6}$$

where

$$\bar{y}_t = \sum_g w_{gt}\bar{y}_{gt} \tag{7}$$

and the weights w_{gt} are the appropriate population or sampling weights at time t.

A problem with this comparison is that, even though there are no changes in the individual group averages $\bar{y}_g$, the $\bar{d}$ may show a significant change simply because the proportion of the population in the various groups is changed, that is, because w_{gt} is not equal to $w_{g(t-1)}$.

A simple and widely used adjustment procedure is to adopt a common set of weights, w_g, to apply to the group means at each period of time [2]. Thus, an adjusted measure of change and adjusted means are given by

$$\bar{d}' = \sum w_g\bar{y}_{gt} - \sum w_g\bar{y}_{g(t-1)} = \sum w_g\bar{d}_g . \tag{8}$$

The common weights to be used may be the sampling or population weights at time t or time (t-1), some average of them, or may be chosen from some external source. Their choice is somewhat arbitrary, but they should be chosen in a rational way for the purpose of the comparison.

We are now ready to consider adjustment for background factors that are multivalued and scaled, or continuous variables, and that may provide unstable background measures over time. For purposes of the exposition, we will assume that there is a simple linear relationship between the outcome y and a background measure x, where it is to be understood that x may be a composite measure constructed by regression or other methods and the result of various linearizing or normalizing transformations. The linear relationship is expressed as

$$y_i' = \mu + \beta x_i + \varepsilon_i \tag{9}$$

where, it is assumed, ε_i are random residuals with mean zero and, at least for the present, we assume x_i is measured without error.

It is convenient to talk about various cases that may arise, based upon assumptions concerning the stability over time of the background variable x and whether or not the regression is common to all groups.

Case 1. The distribution of x has not changed between time (t-1) and time t; in particular, $\bar{X}_{(t-1)} = \bar{X}_t$ where these are the true means at the two dates and

$$\sigma_{X_1} = \sigma_{X_2} .$$

There is common regression across all groups, i.e., $\beta_g = \beta$ and this common regression has not changed over time.

This is a simple case. One can simply adjust the differences in group means at the two times according to the methodology used in regression estimation applied to sample data. Note that

$$\bar{y}_{gt}' = \bar{y}_{gt} + b(\bar{X} - \bar{x}_{gt}) \tag{10}$$

An estimate of the adjusted difference between outcomes for group g between the two dates is as follows:

$$d_g' = \bar{y}_{gt} - \bar{y}_{g(t-1)} + b(\bar{x}_{g(t-1)} - \bar{x}_{gt}) \tag{11}$$

Note that the parameter $\bar{X}$ need not be known since it subtracts out. The adjustment clearly has the effect of reducing sampling error. An aggregate estimate of the difference across all groups is provided by

$$d' = \sum_g w_g d_g' \tag{12}$$

where w_g is chosen as before.

Clearly, one is faced with a dilemma if he is not quite sure that $\bar{X}_{(t-1)} = \bar{X}_t$, $\sigma_{X_1} = \sigma_{X_2}$,

and that the regression coefficients are approximately equal. He can test the hypotheses of equality, but then he becomes involved in the interpretation of sequential tests of hypotheses, and the case is no longer a simple one.

Before proceeding, it may be worth noting that one can adjust for auxiliary variables either by "adjustment by subclassification," as described by Equations (5) and (8), or by regression methods. The method of subclassification is algebraically equivalent to regression when dummy variables (1 or 0) are assigned for each subclass. Cochran [2] has shown that by breaking up "continuous" x variables into classes, one can adjust for major portions of the bias in group comparisons. For monotonic relationships between x and y, his analytical results suggest that one can remove from 64% to 92% of the bias in y by using from two or six classes of x. The method is particularly good when one is uncertain of the relationship between x and y. This may be particularly important in NAEP adjustments since most outcome measures are dichotomous.

It may also be worth noting that, for purposes of adjustment by subclassification, one can tolerate relatively small average frequencies in the adjustment classes since the increased sampling error of small classes is offset by the decreased weight given to each class. (It is only necessary to ensure that an unusually small class does not get a relatively large weight.)

In any case, assuming a regression adjustment for auxiliary variables, whether continuous or not, is simply a convenience in the presentation.

It should also be observed here that a shift in the distribution of the auxiliary variable x can sometimes be adjusted for by a deflator that is external to the survey itself. An obvious example is use of the Consumer Price Index to deflate income. There may also be other deflators that are not commonly used, such as Census estimates of the proportion of persons in specified age groups who have completed high school. It is conceivable that such a deflator could be used to adjust for educational level of parents when t and $t-1$ are widely separated. Also, it is possible that one should concentrate on finding measures of SES that remain relatively stable over time (such as educational attainment of parents) rather than more volatile measures (such as items in the home).

Case 2. The distribution of x has not changed from time 1 to time 2; there is a separate regression within each group that has not changed.

One only needs to replace b by b_g in expressions (10) and (11). There are no further complications.

Case 3. The distribution of x has changed; there is a separate regression within each group that remains constant on the normalized value of x.

Suppose the x variable is income, which can be presumed to change over time. However, it may be that the regression of y on $(x - \bar{X}/\sigma_x)$ will remain constant over time. If so, one can adjust each group mean at time t by

$$d'_g = \bar{y}_{gt} + b_g(\bar{x} - \bar{x}_{gt})/S_x \qquad (13)$$

where S_x is the sample estimate of σ_x and the adjusted gain in outcome can be expressed as

$$d'_g = \bar{y}_{gt} - \bar{y}_{g(t-1)} + b_g(\bar{x}_{g(t-1)} - \bar{x}_{gt})/S_x \qquad (14)$$

In this case, as with the earlier cases, there is the problem of determining whether the regression coefficients have remained fixed and, if they have, how they should be estimated.

What one accomplishes by this adjustment perhaps should be discussed further. Let us suppose that a particular group had an average SES measure at time 1 that fell at the 37th percentile of the national SES distribution. At time 2, their average SES measure may have moved up to (say) the 39th percentile, or (say) down to the 33rd. The adjustment represented by Equation (13) adjusts the average of the outcome measure upward or downward accordingly. Thus, it presumes that the changes in the SES measure are a result of changes in the measurement process, and that equivalent percentile ranks at the two dates identify equivalent SES groups for adjustment purposes.

The interpretation of the adjusted gain is important. It seems evident that, if one's interest lies in evaluating the educational process (both in and out of school), one might very well make such an adjustment because it tends to free the estimate of gain from the gain in the SES measure. However, if one is interested in using educational outcome as a measure of social gain, then it seems inappropriate to make such adjustments. This same principle holds, of course, in all of the adjustments discussed here.

Case 4. The distribution of x has changed; there is a separate regression for each group that has changed and cannot be stabilized by normalization.

This case does not appear to lend itself to adjustment, although some gains might be achieved by assuming one of the simpler models if departures from those assumptions are minor.

There are, of course, other cases, but the ones discussed above appear to be of most interest to NAEP.

5. Additional Comments on Data Adjustment

Much of the interest in adjustment of survey data stems from the desire to infer cause from observed effects in nonexperimental situations. In many social science evaluative studies, it is impossible, within a political system that

recognizes rights of individuals, to experiment with human beings, and often it is unwise or terribly expensive to do so in other cases. Also, even though experimentation might be feasible, the time required for the experiment to run its course may be so great that retrospective surveys are employed. In such cases, one generally tries to accomplish a partitioning of the variation in the outcomes into portions "due to" various characteristics of the observational units, their environments, or the processes to which they have been exposed, or to compare sets of outcomes after such partitioning. (The words "due to" are not to be interpreted as implying cause and effect.)

Sometimes the partitioning of the variation in outcomes is the key analytical result, and a statement such as "Fifty percent of the variation in outcome is accounted for by Factor X" will lead to the conclusion that Factor X needs to be modified through intervention of some kind. Note that cause and effect cannot be inferred from the mathematical statement, but are implied by the decision to intervene. This is a typical exercise in retrospective surveys and has been discussed ably by Dorn [4].

In cases where the analytical result is a comparison of outcomes of two or more groups, focus is usually on controlling bias, i.e., by statistical adjustment for confounding variables. Cochran and Rubin [5] examined, under an assumed linear model, some of the common procedures that have been used, including linear regression adjustment and several matching procedures. Not surprisingly, linear regression adjustment proved to be superior over matching when the linear model with parallel regressions was used and declined in relative merit with respect to "category-matching" with departures from the linear, parallel regression model. Their concept of category-matching is essentially equivalent to our unadjusted and adjusted comparison of groups represented by Equations (5) and (8) above. Their category-matching followed by regression adjustment corresponds closely to procedures we have discussed under Cases 1 through 3.

McKinlay [6] investigated methods for removing bias where the outcome is dichotomous (generally the situation with NAEP data) and the covariate is continuous. Methods investigated were pair-matching and stratification of the covariate. A Monte Carlo analysis of these procedures showed that, for the simulation models studied, pair-matching did not appear to be more effective than stratification. The group comparisons we have discussed in the previous section are quite similar to the concept of stratification on auxiliary variables.

Earlier work, as well as a number of recent studies, have been well summarized by McKinlay [7], and we will not attempt to discuss that work here.

References

1. Morris H. Hansen and Edward C. Bryant, "National Assessment Design Implications," December 1972 meetings of the American Association for the Advancement of Science, Washington, D.C.

2. W.G. Cochran, "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies," Biometrics, 24 (June 1968).

3. Edward C. Bryant, Ezra Glaser, Morris H. Hansen, and Arthur Kirsch, "Associations Between Educational Outcomes and Background Variables: A Review of Selected Literature," under contract to Education Commission of the States, 300 Lincoln Tower, Denver, Colorado, 1974 (Monograph of the National Assessment of Educational Progress).

4. Harold F. Dorn, "Philosophy of Inferences from Retrospective Studies," American Journal of Public Health (June 1953), 677-683.

5. William G. Cochran and Donald B. Rubin, "Controlling Bias in Observational Studies: A Review," available from Donald B. Rubin, Educational Testing Service, Princeton, New Jersey 08540.

6. Sonja M. McKinlay, "Removing Bias Due to a Continuous Covariate from a Dichotomous Response in Pair-Matched and Stratified Samples," Department of Mathematics, Boston University, Boston, Massachusetts.

7. Sonja M. McKinlay, "The Design and Analysis of the Observational Study -- A Review," Harvard University Medical School, Boston Massachusetts (in preparation).