AN EXPERIMENT WITHIN A SURVEY: THE DESIGN OF THE 1990 TRIAL STATE ASSESSMENT OF EDUCATIONAL PROGRESS

Keith Rust and Jim Bethel, Westat, Inc.
1650 Research Blvd., Rockville, MD 20850

1. Introduction

The National Assessment of Educational Progress has monitored the performance of American students at a variety of ages and grades and in a variety of subjects since 1970. In 1990, the National Assessment was extended to include the Trial State Assessment for the purpose of obtaining reliable statistics at the state level.

The 1990 Trial State Assessment of eighth grade mathematics assessed about 2,500 public school students in each of 40 participating states and other jurisdictions. The participating students were assessed in their schools using a one hour pencil and paper test, conducted during February 1990. The results of the assessment were released in June 1991 (see Mullis et al, 1991). The technical details of all aspects of the assessment are provided in Koffler (1991).

One important difference between the national assessment, both historically and as administered in 1990, and the state assessments, was in the mode of administration. The national assessment uses centrally trained personnel temporarily employed by the contractor conducting the administration, and not affiliated with the participating schools and districts. The administration of the state assessments, on the other hand, was the responsibility of the department of education in each participating state, which was free to use school and district personnel to administer the assessment. All administrators were required to attend a one day training session, conducted by the central NAEP administrator (Westat, Inc.).

Because of this difference in administration responsibility, there is concern as to the validity of comparisons between the State and National Assessment data and comparisons between different participating states. In order to assess whether such comparisons were valid, an experimental design was developed where schools were randomly assigned to one of two groups, of which one was monitored by personnel familiar with the National Assessment procedures, while the other group was not monitored. This paper discusses the design of the survey, the experiment to compare the performances of monitored and unmonitored schools, and the methodological issues that were raised by the experiment.

The paper is organized as follows. Section 2 introduces the fundamental aspects of the monitoring procedure, and the rationale for their use. Section 3 discusses the methods used in design for each state. Section 4 details the experimental design used to implement the monitoring procedure, and its relationship to the sample design. In Section 5 procedures for deriving estimates from the assessment are covered. This includes estimates relating to a given state as a whole, and the comparison of the monitored and unmonitored portions of the assessment. This is followed in Section 6 by a discussion of procedures for sample error estimation, via replication, for both types of estimates. In Section 7 the two sets of estimates resulting from the two weighting procedures are compared as to their precision for comparing monitored and unmonitored assessment sessions.

2. Role of the Assessment Monitor

There were two broad areas of concern with regard to the performance of the State Assessment administrators. The first was that the assessment sessions would be conducted in a disorganized manner, thus possibly disadvantaging the participating students. This possibly was made more likely by the complicated nature of the assessment administration (see Koffler (1991)). The second was that there might be intentional efforts to give the students unfair advantage.

The monitoring procedure was developed on the premise that, if the administrator were to make errors due to inadequate training or attention to procedures, this would occur whether or not a monitor were present at the session. If, however, the administrator were inclined to give assistance to one or more students intentionally, this would only take place if no monitor were present. The monitors recorded all deviations from the prescribed procedures, and took corrective action where necessary to insure that a valid assessment resulted. This procedure permitted an evaluation of the effects of various components of inappropriate administration, and gave a subset of the administration (comprising 50 percent of the assessment, as described in the next section) that was known to have been appropriately administered. By recording the errors in administration that they observed, the monitors were able to quantify the extent to which accidental errors in administration occurred throughout the assessment in each state. Comparison of results between monitored and unmonitored sessions provides evidence of whether there was any effect of poor administration or inappropriate assistance.

3. Sample Design

The sample design varied across the forty participants since there were large variations in the numbers of students and schools, and the distribution of students within schools. The common objective was to assess in excess of 2,000 students in each state. The basic model for the design was to draw a stratified sample of about 100 schools, with selection probabilities proportionate to eighth grade enrollment.

Within each selected school an equal probability systematic sample of 30 students was drawn. The objective was to select students from throughout the state with roughly equal selection probabilities, so that there would be at least 2,000 participants after allowance for school and student nonparticipation, and for the exclusion of certain selected students.

In larger states this model was followed exactly. In many states, a few schools were included as certainty selections, with a sample of 30 students being drawn from them as in other schools. In states with fewer than 100 eligible schools, all (in most cases) or most schools were included with certainty, and either 100, 60, or 30 students were selected depending upon the size of the eighth grade enrollment. Special procedures were adopted for those schools, rare in most states but numerous in some, with fewer than 20 eighth grade students. Details of these procedures are given in Bethel et al. (1991).

The stratification variables used were urbanicity, minority enrollment, and household median income of the five-digit zip code area in which the school was located. Urbanicity was the primary stratification variable and consisted of three levels: central city of a Metropolitan Statistical Area (MSA), other MSA, and non-MSA. Minority enrollment strata were derived within urbanicity classes whenever the percentages of black and Hispanic students combined in schools within the urbanicity class exceeded seven percent. Based upon the percentage enrollment of these two minority groups, schools were classified as having high, medium, or low minority enrollment, with one third of the schools from within the urbanicity class belonging to each stratum. An exception was made in urbanicity classes with relatively high enrollment of both black and Hispanic students. Here four strata were formed by simultaneously classifying schools by whether they had relatively high or low black enrollment, and whether they had relatively high or low Hispanic enrollment. Within the explicit strata schools were sorted by the income variable described above. A systematic sample of schools was drawn by sorting schools by urbanicity class, minority stratum, and income. The sorts by minority stratum and income were such that successive schools on the list were relatively similar on at least two of the three selection variables.

Within selected schools, the eighth grade students were listed by the school. The most commonly used listing order was alphabetic. A systematic random sample of thirty students was drawn from the list.

There were several other aspects of the sampling procedure that will not be discussed. These include the supplemental sampling of newly opened schools, substitution procedures for nonparticipating schools, rules for the exclusion of students, and supplemental samples of newly enrolled students. These features are discussed in Bethel et al. (1991).

4. Assignment of Schools to be Monitored

It was decided that fifty percent of the schools within each state were to have their assessment sessions monitored. The exceptions were the two territories, with only six schools each, where half of the sessions in each school (of which there were up to ten) were monitored. Fifty percent monitoring was used to maximize the power to detect a difference between the performance of students in monitored sessions and that of those in unmonitored sessions.

The assignment was achieved by pairing schools based on the ordering used in sample selection. Thus the members of each pair were similar with regard to the school stratification characteristics. Certainty schools were listed separately, using an ordering derived from the values of the stratification variables. One member of each pair was chosen with probability of 0.5 to be monitored, with the assignment being independent from pair to pair. Thus the sample design can be viewed as consisting of 50 strata, with one selection per stratum being made for the monitored sample, and one per stratum for the unmonitored sample, with no overlap between the two samples. This procedure meant that the two half samples were as similar as possible, subject to the practical constraint that all students from the same school had to be assigned the same monitor status (monitored or unmonitored).

In the event, the monitoring occurred almost as assigned. All schools assigned to be unmonitored were in fact unmonitored, and only in occasional cases was it not possible for the monitor to keep the appointment to attend the session. Overall, about 49.1 percent of participants attended a monitored session.

5. Survey Estimation

The assessment results for the Trial State Program were derived using survey weighting procedures which account for selection probabilities and school and student non-participation. No account was taken explicitly of the monitor status of the students. The results released for each state are the simple aggregate of the data from monitored and unmonitored students. This approach was predicated upon three facts: first, there was little evidence of difference between monitored and unmonitored sessions; second, the monitoring was conducted at the 50 percent level in each state across the various school survey strata; third, precision of estimation would be enhanced by not introducing variation in the weights to account for random differences in the composition of the two half samples.

In brief, the weight for each student consisted of four components. For student j from school i, W_i denotes the inverse of the selection probability of the school. This weight was adjusted by a nonresponse adjustment factor, f_c, where c denotes the school weighting class to which school i belongs. The third component is the within school skip factor used to draw the systematic sample of students, s_i. Finally, a set of student nonresponse adjustment factors, f_{2a}, constant for students belonging to student weighting class a, was derived. The weight for student j from school i is given by

$$W_{ij} = W_i f_c s_i f_{2a}$$

where student j belongs to class a and school i belongs to class c. Estimates of the distributions of student proficiency, and multivariate analysis of student level data are performed by weighting the record for student j from school i by W_{ij}.

A separate set of weights was developed for use in comparing the results of monitored and unmonitored sessions. These weights were designed to give approximately unbiased separate estimates of the proficiencies of monitored and unmonitored students. The full sample weights described above have this property also; the special "comparison" weights are intended to give estimates of the differences between the performance of monitored and unmonitored students that have less sampling variance than those obtained from the full sample weights. In Section 7 empirical comparisons are presented showing the results from these two procedures.

The comparison weights were derived from the full sample weights. Adjustments were made to the weights so that the marginal distributions for various student characteristics would be equal for the two half samples. This meant that random differences in these distributions, based on the full sample weights, would not contribute to the variance of the half sample comparisons. The marginal distributions used for this procedure were for characteristics not controlled by the school stratification, and that were correlated with eighth grade mathematics proficiency. Examination of the results of the 1986 NAEP mathematics assessment at grade 7 suggested six variables that could be utilized. These variables are student level variables, listed below, and thus there was substantial scope for variation between the two half samples as to their distribution.

In order to equalize the weighted distributions of the 1,200 or so monitored and unmonitored student in each state, without introducing undue variability into the weights, and to reflect some important interactions between the six variables listed, the following approach was used. For the full sample in each state, three marginal distributions were formed: sex by age, race/ethnicity by parents' education; "type of math course taken" by "attitude to mathematics". The marginal cells were collapsed as necessary to ensure that there were at least seventy-five sampled students in each such cell.

Next, a four step ranking procedure (Oh and Scheuren, 1987) was used to adjust the full sample weights for each half sample to agree with the full sample weights for the whole sample, with respect to each of the three marginal distributions. The weights were first raked to the sex by age marginal, then the race/ethnicity by parents' education marginal, then the type of math course by math attitude marginal, and finally the sex by age marginal again. A preliminary study showed that no gains of any significance would accrue from further raking, since all three marginal distributions were closely matched following these four steps.

Let the full sample weight for student ℓ from class i of the first marginal, class j of the second, class k of the third and monitor status m, be denoted as $W_{ijk\ell m}$.

Let $$N_{i..} = \sum_m \sum_j \sum_k \sum_\ell W_{ijk\ell m}$$

With $N_{.j.}$ and $N_{..k}$ defined analogously.

Let $$\tilde{N}^{(0)}_{mijk} = \sum_\ell W_{ijk\ell m},$$

$$\tilde{N}^{(a+1)}_{mijk} = \frac{\tilde{N}^{(a)}_{mijk} N_{i..}}{\sum_j \sum_k \tilde{N}^{(a)}_{mijk}} \quad \text{for } a = 0,3.$$

$$\tilde{N}^{(2)}_{mijk} = \tilde{N}^{(1)}_{mijk} N_{.j.} / \sum_i \sum_k \tilde{N}^{(1)}_{mijk}$$

$$\tilde{N}^{(3)}_{mijk} = \tilde{N}^{(2)}_{mijk} N_{..k} / \sum_i \sum_j \tilde{N}^{(2)}_{ijk}$$

The comparison weight for each student is given by

$$W^*_{mijk\ell} = W_{mijk\ell} \tilde{N}^{(4)}_{mijk} / \tilde{N}^{(0)}_{mijk}.$$

The final weight for each student did not vary wildly from the initial full sample weight following this procedure. Apart from the factor of two resulting from weighting the half sample to the full sample distribution of weights, the variations for individual students were mostly in the range of 0.5 to 2.0, with very few outside the range of 0.33 to 3.0. This held consistently across states.

6. Replicated Variance Estimation

Variances for NAEP are estimated using the Jackknife Repeated Replication procedure, based on an approximation of the sample design as having two first stage units drawn with replacement from each explicit stratum (see Wolter (1985, Chapter 4)). The replicated variance estimates are obtained by using a set of replicate weights (see Dippo, Fay and Morganstein (1984)). Two sets of replicate weights were produced, one corresponding to the full sample weights and the other corresponding to the comparison weights. These separate sets of replicate weights were needed for two reasons. The first, which applied in every state, was to account for the effect on sampling variance of the raking weight adjustments applied to the comparison weights, but not the full sample weights, as described in the previous section. The second reason applied only in states having one or more certainty school selections. These certainty selections were handled differently for estimating variances for the aggregate of the monitored and unmonitored samples, and for comparisons between these half samples. Details of this difference in procedure, and the reasons for it, are given below.

The procedure for replication for the full sample was as follows. In order to replicate the design, it was necessary to form pairs of non-certainty schools, with the schools in each being similar with respect to stratification variables, and of the same **assigned** monitor status. Failure to do this would have led to overestimation of variance, since then variability among the replicate estimates would have resulted from differences in stratification characteristics of the school, and any differences resulting from the impact of monitoring. The stratification procedure used, and the random assignment of exactly 50 percent monitoring of schools within "strata", meant that these sources of variance were in fact negligible.

Thus non-certainty schools were sorted in the order in which they were selected, and each successive pair of monitored schools constituted a subset of replicate pairs, with the rest being formed from successive pairs of unmonitored schools. For certainty schools, most often a single school constituted one replicate pair, the students being randomly assigned half each to each member of the pair. For some states particularly large schools each constituted two pairs. For a given replicate, the base weights W_i was set to zero for one member of a given replicate pair, the member being chosen at random. The weight W_i for the remaining pair member was doubled. The school and student nonresponse adjustments were then recalculated for the whole sample, utilizing these perturbed replicate base weights, and a final replicate weight was composed from these adjustments and the replicate base weights. The full set of replicate weights was formed by repeating this process using each replicate pair in turn. About 50 replicate weights per student were formed in this way in each state.

To conceptualize the reasoning for not reflecting a variance component due to the assignment of monitor status consider the following hierarchical linear model for mathematics proficiency. For simplicity, the impact of stratification will be ignored. The mathematics proficiency x for student j from school i with monitor status m is given by

$$x_{mij} = \mu + \alpha_m + \beta_{i(m)} + \varepsilon_{j(mi)}$$

where μ denotes the overall mean, α_m denotes the effect of monitor status m, $\beta_{i(m)}$ denotes the effect of school i nested within monitor status m, and $\varepsilon_{j(mi)}$ denotes the effect of student j, nested within school i and monitor status m. The monitoring effect is fixed, with $\sum_{m=0}^{1}\alpha_m=0$, and the student effect is random with $E(\varepsilon_{j(mi)}) = 0$, and $E(\varepsilon^2_{j(mi)}) = \sigma^2_\varepsilon$. For non-certainty schools, the school effect is random with $E(\beta_{i(m)}) = 0$ and $E(\beta^2_{i(m)}) = \sigma^2_\beta$, while for certainty schools the $\beta_{i(m)}$ are fixed effects, with $\sum_m \sum_i \beta_{i(m)} = 0$.

Consider a two-stage sample drawn with replacement at each stage and equal probability overall in which 2I schools are selected and half are assigned to be monitored. Within each school n students are selected. Consider the estimate of mean $\bar{x}$, where

$$\bar{x} = \frac{1}{2In}\sum_{m=0}^{1}\sum_{i=1}^{I}\sum_{j=1}^{n} x_{mij}.$$

Using the conditional variance approach (Cochran (1977, Chapter 10)),

$$V(\bar{x}) = E_1 V_2(\bar{x}) + V_1 E_2(\bar{x}).$$

where the subscript 2 denotes "conditional on the selection of schools", and subscript 1 denotes "with respect to the school sampling procedure". It follows from the model that

$$E_1 V_2(\bar{x}) = V_2(\bar{x}) = \sigma^2_\varepsilon / (2In)$$

$$E_2(\bar{x}) = \mu + \frac{1}{2I}\sum_m \sum_i \beta_{i(m)}.$$

Hence

$$V_1 E_2(\bar{x}) = \sigma^2_\beta / 2I \text{ for non-certainty schools}$$

$$= 0 \text{ for certainty schools.}$$

Thus

$$V(\bar{x}) = \frac{1}{2I}(\sigma^2_\beta + \sigma^2_\varepsilon / n) \text{ for non-certainty schools}$$

$$= \sigma^2_\varepsilon / 2In \text{ for certainty schools.}$$

Thus the effect of monitoring has no impact upon the variance of $\bar{x}$, whereas the effect of student sampling within schools and, in the case of non-certainty schools, the effect of school sampling within strata, do need to be reflected. The replication scheme utilized appropriately reflects these various variance components.

Two things are important to note about this model and its implications for variance estimation. The first is that, while allowing for an effect of monitoring (which can in fact vary across strata) there is no within stratum interaction term between school and monitoring status. With the hierarchical design used, it is not possible to estimate such an interaction. The variance estimation scheme thus assumes that this interaction is zero (but note that this is not a source of bias in the estimates themselves since monitoring status was assigned randomly). The alternative would have been to overestimate the variance by reflecting a component of variance for the main effect of monitoring, which through the design was effectively zero. The second feature is that for certainty schools in general $\sum_i \beta_{i(m)} \neq 0$ for each m. This has implications for the replicated variance estimation for comparisons between monitored and unmonitored assessments.

Consider the comparison between the monitored and unmonitored half samples within a given state. With

$$\bar{x}_m = \frac{1}{In} \sum_{i=1}^{I} \sum_{j=1}^{n} x_{mij}$$

the estimate of difference in proficiency under monitoring compared with no monitoring is given by

$$\hat{d} = \bar{x}_1 - \bar{x}_0.$$

Using the model and the same approach as above

$$V(\hat{d}) = E_1 V_2(\hat{d}) + V_1 E_2(\hat{d}).$$

It follows easily that

$$E_1 V_2(\hat{d}) = V_2(\hat{d}) = 2\sigma_\varepsilon^2 / (In),$$

while

$$E_2(\hat{d}) = (\alpha_1 - \alpha_0) + \frac{1}{I}\left(\sum_{i=1}^{I} \beta_{i(1)} - \sum_{i=1}^{I} \beta_{i(0)}\right).$$

Note that for certainty schools, this does not in general reduce to $(\alpha_1 - \alpha_0)$. In fact

$$V_1 E_2(\hat{d}) = 2\sigma_\beta^2 / I$$

for both non-certainty schools and certainty schools, ignoring finite population arguments. (Note that, even if a finite population correction were to be included, for certainty schools this term does involve a component due to variation among schools.)

Thus

$$V(\hat{d}) = \frac{2}{I}(\sigma_\beta^2 + \sigma_\varepsilon^2 / n).$$

Comparing this with the formula for $V(\bar{x})$ shows that the same procedure for forming replicate pairs can be used to analyze $\hat{d}$ and $\bar{x}$ in the case of non-certainty schools, but that this is not so for certainty schools.

Thus a second set of replicate weights was formed for use in comparing monitored and unmonitored sessions. Replicates for certainty schools in this case were formed by pairing successive monitored certainty selections, and successive unmonitored certainties, with the ordering being based on the values of the variables used in stratification. Replicates for non-certainty schools were formed in the same manner as for the full sample replicates discussed above.

In addition, as mentioned the raking adjustments to the weights were replicated in forming replicate weights for estimating the variances of comparisons. That is, once the base weights W_i were appropriately perturbed for a given replicate, the school and student nonresponse adjustments were recomputed using these, and then the raking procedure was applied to each replicate. This meant that the raking procedure was repeated a total of over 50 times in most states. For this reason it was important to have a small fixed number of iterations for raking rather than specifying a convergence criterion, since using the latter would have proved prohibitively expensive.

7. Comparison of Weighting Procedures

As discussed in Section 5, two sets of weights were developed for the assessment. Both are designed to provide approximately unbiased estimation. The one set is appropriate for estimating characteristics and proficiencies of students without regard to the use of monitoring, while the second set is designed to give greater precision for comparisons between monitored and unmonitored assessments. For each set of weights a set of replicate weights has been developed to provide approximately unbiased estimates of the appropriate sampling error associated with using the particular set of estimation weights.

In this section we provide an evaluation of whether in fact this second set of weights gave greater precision for comparing monitored and unmonitored sessions. We compare the estimates of variance for comparisons between monitored and unmonitored students. We consider the results in those states with no certainty schools since, as discussed in Section 6, the presence of certainty schools contributes variance to the comparisons of monitored and unmonitored students, but not to aggregate estimates. Table 1 shows a series of estimates and their standard errors for each state. The same quantity is estimated using each set of weights, so that the standard errors can be compared to see which set of weights gives greater precision. Estimates for the whole state and for demographic subgroups are presented.

The results show that, as would be expected from the linear model development in Section 6, the standard errors for comparisons using overall weights are about two times those of the overall mean estimate. Using the special comparison weights and their associated replicate weights reduces these standard error substantially, by about 40 percent or so for the whole-state estimates. Clearly, the use of the raking procedure, and of inference conditional on the marginal distributions of various student characteristics, has reduced the level of sampling error substantially, notwithstanding the variability introduced to the weights by the raking procedure. Those gains realized appear to occur primarily for the whole population and for subclasses that are well distributed across schools (e.g., males), and are only minor for subclasses that tend to be clustered by school (e.g., Blacks).

Acknowledgements

The major components for the design of the 1990 Trial State Assessment were developed during 1988 in a series of meetings of NAEP personnel from the National Center for Education Statistics, Educational Testing Service and Westat. Although the design was a group effort, the late Morris Hansen was a major

architect of the design and initiated much of the development for a plan to monitor the assessment in each state.

Our thanks go to Gene Johnson and Bruce Kaplan at ETS for providing the estimates in Table 1.

The views expressed in this manuscript are those of the authors alone, and do not necessarily reflect the opinions of any of the organizations involved in the NAEP project.

References

Bethel, J., Rust, K., and Severynse, J. (1991). Chapter 3 in Koffler, S., The Technical Report of NAEP's 1990 Trial State Assessment Program. Educational Testing Service: Princeton, NJ.

Cochran, W.G. (1977). Sampling Techniques. 3rd Edition. Wiley: New York.

Dippo, C.S., Fay, R.E., and Morganstein, D. (1984). Computing Variances from Complex Samples with Replicate Weights. Proceedings of the Section on Survey Research Methods. American Statistical Association: Washington, D.C.

Koffler, S. (1991). The Technical Report of NAEP's 1990 Trial State Assessment Program. Educational Testing Service: Princeton, NJ.

Mullis, I.V.S., Dossey, J.A., Owen, E.H., and Phillips, G.W. (1991). The State of Mathematics Achievement. National Center for Education Statistics: Washington, DC.

Oh, H.L., and Scheuren, F. (1987). Modified Raking Ratio Estimation. Survey Methodology, Vol. 13, pp. 209-219.

Wolter, K.M. (1985). Introduction to Variance Estimation. Springer-Verlag: New York.

Table 1. Sampling errors for means and monitored-nonmonitored comparisons, using overall weights and comparison weights

	Overall proficiency mean		Monitored-nonmonitored comparison			
	Overall weights		Overall weights		Comparison weights	
State	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
California	256.2	1.3	-8.6	2.5	-5.5	1.5
Florida	254.8	1.2	+4.6	2.5	+3.7	1.7
Georgia	258.2	1.3	-0.6	2.6	-0.9	1.4
Illinois	260.4	1.7	-6.4	3.6	-5.4	2.0
Michigan	264.2	1.1	+2.7	2.6	+1.3	1.3
Ohio	263.8	1.0	-5.1	2.4	-2.2	1.4
New York	260.7	1.3	+3.7	3.0	+1.5	1.7
North Carolina	249.6	0.9	-4.6	2.0	-0.9	1.2
Pennsylvania	266.3	1.6	-2.1	2.9	-1.7	1.4
Texas	257.8	1.3	-0.7	2.6	+1.0	1.6
Males						
California	257.5	1.6	-9.2	3.2	-6.5	2.2
Florida	256.5	1.6	+5.0	3.2	+3.7	2.2
Georgia	258.6	1.6	-2.0	3.2	-2.1	2.1
Illinois	260.4	1.7	-5.1	3.9	-3.5	2.5
Michigan	265.2	1.3	+5.5	3.1	+3.2	1.7
Ohio	266.5	1.2	-5.3	2.6	-1.9	1.9
New York	262.5	1.5	+7.5	3.2	+5.1	2.2
North Carolina	249.0	1.2	-4.8	2.4	-1.0	1.7
Pennsylvania	269.0	1.7	-0.6	3.3	-0.4	1.9
Texas	259.6	1.5	+2.6	3.0	+4.2	2.0
Blacks						
California	233.4	2.9	-0.8	5.8	+0.6	5.4
Florida	231.0	1.7	+1.0	3.4	+1.3	3.2
Georgia	239.2	1.3	-7.6	2.7	-4.8	2.6
Illinois	232.6	3.8	-4.8	7.0	-4.4	5.5
Michigan	230.2	1.4	+3.1	3.1	+2.6	3.2
Ohio	232.4	1.2	-4.8	3.3	-3.9	3.2
New York	236.5	2.5	+2.9	5.4	+2.5	4.7
North Carolina	231.7	1.1	-1.7	2.1	-0.5	2.2
Pennsylvania	237.5	3.4	+9.9	7.5	+8.8	6.9
Texas	234.0	1.6	-5.8	3.3	-5.2	3.5