National Center for Education Statistics

U.S. Department of Education
Institute of Education Sciences
NCES 2005–01

Working Paper Series

2000 NAEP — 1999 TIMSS Linking Report

April 2005

The Working Paper Series was initiated to promote the sharing of the valuable work experience and knowledge reflected in these preliminary reports. These reports are viewed as works in progress, and have not undergone a rigorous review for consistency with NCES Statistical Standards prior to inclusion in the Working Paper Series.

U.S. Department of Education
Margaret Spellings
Secretary

Institute of Education Sciences
Grover J. Whitehurst
Director

National Center for Education Statistics
Grover J. Whitehurst
Acting Commissioner

The National Center for Education Statistics (NCES) is the primary federal entity for collecting, analyzing, and reporting data related to education in the United States and other nations. It fulfills a congressional mandate to collect, collate, analyze, and report full and complete statistics on the condition of education in the United States; conduct and publish reports and specialized analyses of the meaning and significance of such statistics; assist state and local education agencies in improving their statistical systems; and review and report on education activities in foreign countries.

NCES activities are designed to address high priority education data needs; provide consistent, reliable, complete, and accurate indicators of education status and trends; and report timely, useful, and high quality data to the U.S. Department of Education, the Congress, the states, other education policymakers, practitioners, data users, and the general public.

We strive to make our products available in a variety of formats and in language that is appropriate to a variety of audiences. You, as our customer, are the best judge of our success in communicating information effectively. If you have any comments or suggestions about this or any other NCES product or report, we would like to hear from you. Please direct your comments to:

National Center for Education Statistics
Institute of Education Sciences
U.S. Department of Education
1990 K Street NW
Washington, DC 20006–5651

April 2005

The NCES World Wide Web Home Page address is *http://nces.ed.gov*
The NCES World Wide Web Electronic Catalog is: *http://nces.ed.gov/pubsearch*

Suggested Citation

Johnson, E., Cohen, J., Chen, W.H., Jiang, T., and Zhang, Y. (2003). *2000 NAEP—1999 TIMSS Linking Report* (NCES 2005–01). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Content Contact:
Steven Gorman
NAEP Program Director Assessment Design & Analysis
(202)502-7347
Steven.Gorman@ed.gov

Working Paper Foreword

In addition to official NCES publications, NCES staff and individuals commissioned by NCES produce preliminary research reports that include analyses of survey results, and presentations of technical, methodological, and statistical evaluation issues.

The ***Working Paper Series*** was initiated to promote the sharing of the valuable work experience and knowledge reflected in these preliminary reports. These reports are viewed as works in progress, and have not undergone a rigorous review for consistency with NCES Statistical Standards prior to inclusion in the Working Paper Series.

Copies of Working Papers can be downloaded as pdf files from the NCES Electronic Catalog (http://nces.ed.gov/pubsearch/).

Marilyn M. Seastrom
Chief Mathematical Statistician
Statistical Standards Program

Ralph Lee
Mathematical Statistician
Statistical Standards Program

2000 NAEP — 1999 TIMSS LINKING REPORT

March 31, 2003

Prepared by:

The American Institutes for Research
The Educational Testing Service

Authored by:

Eugene Johnson
Jon Cohen
Wen-Hung Chen
Tao Jiang
Yu Zhang

With Contributions by:

John Mazzeo
Laura Jerry
Ed Kulick
Satwinder Thind

TABLE OF CONTENTS

CHAPTER I
INTRODUCTION

This is the second study linking NAEP to TIMSS. The first study linked the 1996 NAEP to the 1995 TIMSS (Johnson, 1998). This study attempted to link the 2000 grade 8 NAEP in mathematics and science to the 1999 grade 8 TIMSS (which also assessed mathematics and science). The major purpose of both studies, assuming a successful link, was to allow comparisons of states that participated in NAEP with nations that participated in TIMSS.

The earlier study offered little opportunity for validation of the resulting linkage, and the possible validations yielded mixed results. The link worked at grade 8 in the sense that the predicted TIMSS results for Minnesota and Minnesota's actual TIMSS results were close to each other. The link did not work at grade 4. The predicted TIMSS results for both Colorado and Minnesota were considerably higher than the actual TIMSS results. What went wrong with the grade 4 linkage was never definitively determined.

One of the problems in the first linking was that no student took both assessments. This means that any linking depends on the assumption of equivalent populations and the untestable assumption that a given level of performance on one assessment implies a certain performance on the other. This time, we have a set of data, the *linking sample*, where students responded to both NAEP and TIMSS using the administration conditions appropriate to each assessment. This allows a direct comparison of performance on the two instruments, permits much stronger statistical linking models, and removes all doubt about the comparability of the people taking each instrument. Roughly 2,000 students who took the mathematics NAEP also took the TIMSS[1]. Similarly, roughly 2,000 students who took the science NAEP also took the TIMSS.[2]

[1] Each TIMSS booklet included both mathematics and science items. Only the TIMSS mathematics scores were used in any linkings involving the students also taking NAEP mathematics.
[2] Only the TIMSS science scores were used in any linkings involving the students also taking NAEP science.

A projection-based linkage forms the core of this study. These linkages were based on regression functions applied to the NAEP and TIMSS data from the linking sample. We then used estimates from these regressions to project state-average TIMSS scores from NAEP scores in the 12 states that participated in both studies and compare the projected TIMSS scores with actual TIMSS scores. As this report shows, this linkage did not fare well—the linkage significantly underpredicted actual TIMSS scores. This result suggests that TIMSS functioned differently in the linking sample than in the national sample. For this reason, no predicted state TIMSS scores were created for the states participating in NAEP but not in TIMSS.

We confirmed the differential functioning of TIMSS in the linking sample by developing a linkage based on the national data and projecting it to the linking sample. In the national data, no students took both TIMSS and NAEP, so the link had to be made by matching the means and standard deviations of the respective reported distributions. This *moderation linkage* followed the methodology of the 1996 NAEP/1995 TIMSS link. As in that link, we evaluated the stability of the linking function by examining its stability for demographic subgroups defined by race/ethnicity, gender, and other demographic characteristics. We estimated components of variance of the link owing to model misfit, measurement error, sampling error, and instability over subgroups.

Using this moderation linkage, we projected TIMSS scores from NAEP onto the linking sample. This linkage overpredicted the observed TIMSS scores, confirming the finding from the primary (projection) linkage. With this we were confident that performance on TIMSS differed between the linking sample and the national sample. We dedicated the remainder of the study to an investigation of the factors that may have caused this discrepancy.

It should be noted that the moderation linkage did a good job of predicting the TIMSS scores for those states that participated in both NAEP and TIMSS. Since this linkage was based the national data for NAEP and TIMSS, problems due to the differential functioning of TIMSS in the linking sample are not present in the moderation linkage.

Reporting on these methods and results, this report proceeds in four more chapters.

- Chapter 2, *NAEP and TIMSS and the Linking Sample*, describes the national NAEP and TIMSS samples as well as the linking sample. Chapter 2 reveals some basic discrepancies in performance on the two assessments in the linking sample.
- Chapter 3, *Projection Linking*, describes the methods used to estimate the linking function via projection, along with the methods used to estimate the precision of these estimates. Estimated parameters of the projection linking function and associated standard errors are reported there.
- Chapter 4, *Validation and Investigations*, evaluates the validity of the linkage, with the conclusion that TIMSS performed differently in the linking sample than in the national study. Chapter 4 presents the findings of a set of studies designed to confirm this finding including the results of a moderation linkage based on the national data, evaluates motivational differences between the linking and national samples, and reviews a study of the content differences between NAEP and TIMSS.
- Chapter 5, *Discussion*, summarizes the findings and explores their implications for NAEP, TIMSS, the design of linking studies, and testing in general.

CHAPTER II
NAEP AND TIMSS AND THE LINKING SAMPLE

NAEP AND TIMSS

NAEP is an ongoing, Congressionally mandated survey designed to measure what students know and can do. The goal of NAEP is to estimate educational achievement and changes in that achievement over time for American students of specified grades as well as for subpopulations defined by demographic characteristics and by specific background characteristics and experiences. In 2000, NAEP collected mathematics and science data from nationally representative samples of students in public and private schools in grades 4, 8, and 12. Additionally, directly comparable state assessments were conducted in public schools in participating states and jurisdictions at grade 4 for mathematics and at grade 8 for mathematics and science. Earlier State NAEP also collected information about private schools. However, state-level NAEP mathematics and science results are available for grade 8 public school students in 44 states and jurisdictions. Further details about the 2000 NAEP can be found in the 2000 NAEP Technical Report.

TIMSS began as the Third International Mathematics and Science Study, the largest and most ambitious study ever conducted by the International Association for the Evaluation of Educational Achievement (IEA). In 1999 the assessment was repeated and renamed the *Trends* in International Mathematics and Science Study (and referred to as TIMSS-99 for the 1999 assessment). TIMSS remains an international comparative study designed to provide information about educational achievement and learning contexts for the participating countries. In 1999, 38 countries collected TIMSS data in 34 languages. Each participating country assessed mathematics and science in the two grades with the largest proportion of 13-year-olds (grades 7 and 8 in most countries, including the United States). Mathematics and science results are available for 38 countries for the higher of these grade levels-which, for convenience, we refer to as the grade 8 level in this report. Further details

about the 1999 TIMSS can be found in the 1999 TIMSS Technical Report.

The U.S. results are based on a sample of students from public and private schools. In addition, 13 states participated in a state-level administration of grade 8 TIMSS mathematics and science in 1999. Twelve of those states also participated in the 2000 State NAEP. Thus, released public school NAEP and TIMSS results are available for those 12 states.

A number of key characteristics of the NAEP and TIMSS results have a bearing on the adequacy of any link between the two assessments. These include the following:

- Both NAEP and TIMSS are based on complex probability samples of the student population. Both U.S. samples include public and private students in grade 8. The sample sizes for the two assessments in the United States are similar: 15,694 and 9,072 for NAEP and TIMSS grade 8 mathematics, and 15,787 and 9,072 for NAEP and TIMSS grade 8 science, respectively.

- TIMSS was conducted in the United States (and in most Northern Hemisphere countries) in February to May of 1999. NAEP was conducted January through March 2000. Thus, the TIMSS results are applicable to the achievement of the 1999 student population at the end of the school year, whereas the NAEP results are applicable to the achievement of the 2000 student population some months before the end of the school year.

- The frameworks that defined the NAEP and TIMSS assessments are not identical but appear similar. Both assessments include multiple-choice and short and extended constructed-response questions, but NAEP has a higher proportion of constructed response items than does TIMSS. The two assessments have no items in common.

- In TIMSS, the same students participated in both the mathematics and science testing, with 90 minutes total testing time across the two subjects or 45 minutes for each. Both mathematics and science were mixed in each booklet. In NAEP, each sampled student received either a mathematics or a science instrument. Total testing time for the mathematics instrument was 45 minutes, comparable to TIMSS mathematics. Total testing time for NAEP science was 90 minutes at grade 8, including 30 minutes of hands-on tasks.

- Students participating in NAEP are selected randomly from within schools (independent of their classroom assignments) and are removed from their classrooms for testing. TIMSS tests students in intact classrooms, assessing all students in selected classrooms.

- Both NAEP and TIMSS scaled their data using Item Response Theory (IRT) techniques. Both studies used a variety of scaling models (two and three parameter logistic and generalized partial credit) to develop subscales for mathematics and science. NAEP mathematics and science composites were then created as weighted averages of the

mathematics and science subscales. TIMSS calibrated all items together to create separate overall scales, one for mathematics, and one for science.

- Both NAEP and TIMSS used a methodology to account for the imprecision of measurement of individual students abilities (plausible values). These allow appropriate estimates for any subgroups contained in the conditioning model.

Clearly, although they are similar, the NAEP and TIMSS assessments do differ in ways that will impact the link between the two.

THE LINKING SAMPLE

As mentioned in the introduction, TIMSS was administered to a subsample of NAEP respondents. This set of students responded to both assessments, facilitating an examinee-level linkage. We refer to this sample of students who responded to both instruments as the *linking sample.*

Selection and size

The linking sample is a convenience sample rather than a probability sample. The sample includes schools that participated in NAEP mathematics or science assessments in 2000 and volunteered to participate in the linking study. Schools were paid a fee as an incentive to volunteer. The pool of volunteers was monitored to ensure diversity in SES, community size, and region. The sample includes schools from 46 states or jurisdictions. The final mathematics linking sample includes 1,741 students from 86 schools, and the science linking sample includes 1,818 students from 89 schools. [1]

Student scores in the linking sample

Every examinee in the linking sample was part of the operational NAEP national sample in 2000. Hence, NAEP operational plausible values were available for each linking-sample examinee.

Plausible values are discussed in Chapter 3, and more detail is available in the 2000 NAEP technical report. The NAEP plausible values provided the measures of NAEP performance used in the linking study.

TIMSS performance was directly estimated from item responses and the TIMSS operational IRT item parameters (see Chapter 3 for more discussion of this method).[2]

Student performance in the linking sample

Since the linking samples are not random, the average in the linking sample might depart from the national sample average. Table 2.1 compares the average TIMSS and NAEP scores in the linking samples with the corresponding scores in the national samples.

Table 2.1: Average TIMSS and NAEP mathematics and science scores in the linking and national samples

	TIMSS			NAEP			Linking Sample Size
	Linking Sample	1999 National Sample	Difference	Linking Sample	2000 National Sample	Difference	N
Mathematics							
All Students	479.9	501.6	-21.7	274.1	275.5	-1.4	1741
Public School	475.5	498.2	-22.7	272.3	274.4	-2.1	987
Science							

[1] The mathematics linking sample includes 987 students from public schools. The science linking sample includes 976 students from public schools.

[2] We also estimated the overall linking function by using plausible values specially designed for the linking function. These TIMSS-linking plausible values were conditioned on NAEP posterior means, along with a several demographic variables and key interaction terms. The plausible values yielded virtually identical results for the main linking functions. "Direct estimation" using a modified version of the original marginal maximum likelihood regression module in AM (Cohen and Jiang, 2001) provides a confirmation of the plausible values-based results and offered a convenient way to provide a means for aggregating sources of error. Therefore, the results reported here are based on direct estimates. It should be noted that the linking sample values used the post-stratified weights given in Appendix A.

All Students	497.9	515.0	-17.1	153.2	154.1	-0.9	1818
Public School	493.0	510.4	-17.4	151.6	149.2	2.4	976

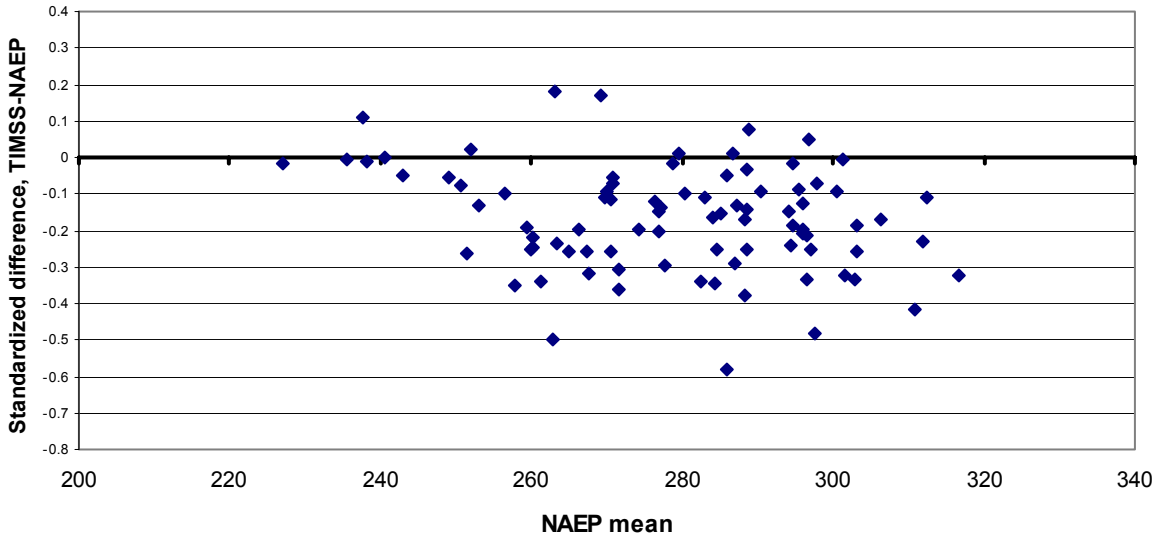
SOURCES: National tabulations based on data from the 1999 IEA Trends in International Mathematics and Science Study (TIMSS-99) and the 2000 National Assessment of Educational Progress (NAEP). Linking sample tabulations based on a special sample of students receiving both NAEP and TIMSS in 2000.

The average scores for NAEP mathematics and science of the linking sample are close to the national average regardless of the school type, 1.4 to 2.1 points below for mathematics and 0.9 points below for all students in science and 2.4 points above for public school students in science. A larger discrepancy in the average score appears in TIMSS between the linking and national samples. The schools in the linking sample scored more than 20 points below the national mean in TIMSS for mathematics and around 17 points below in science. Owing to the differences in the reporting scales, 1 TIMSS point is roughly equivalent to about 2 NAEP points in mathematics and 3 NAEP points in science. Thus, the discrepancies in TIMSS mathematics are about 12 to 13 NAEP points compared with about 2 points in NAEP. The discrepancies in TIMSS science are about 5 NAEP points compared with –0.9 to 2 points in NAEP.

The discrepancy in performance in the linking sample—about average on NAEP and below average on TIMSS—appeared consistently across schools. We can see this in Figures 2.1 and 2.2. These figures display the school-level difference between TIMSS and NAEP scores in a standardized metric, and plot them by NAEP score. For each figure, scores on both instruments have been standardized so that a score of zero corresponds to the *national* mean, and the standard deviation has been standardized so that the *national* standard deviation is 1.[3] The vertical axis plots the difference between these standardized scores, and the horizontal axis plots the school's average NAEP score.

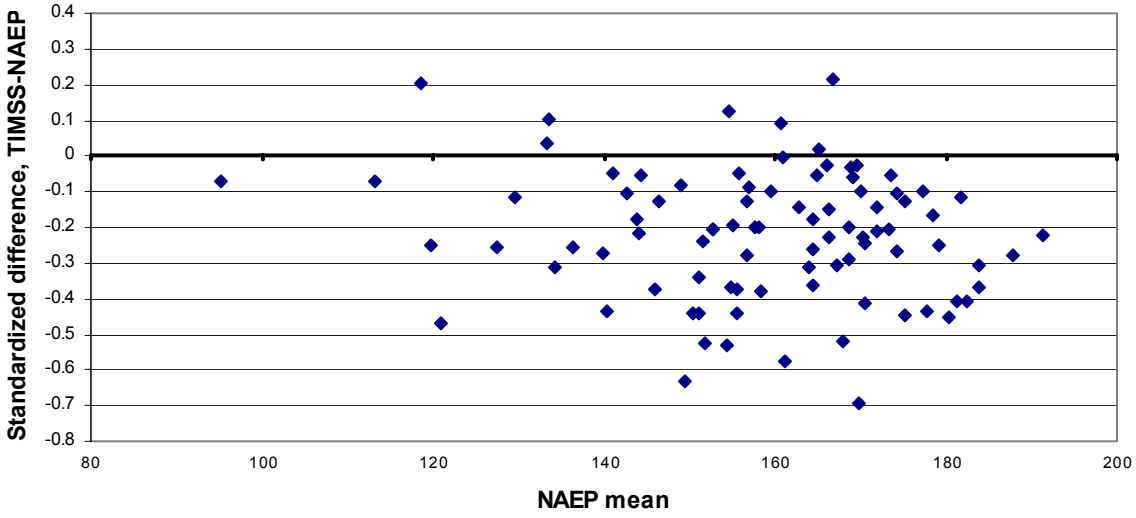
[3] These figures are based on the first of the plausible values generated for the linking study. Direct estimates and averages of plausible values tell the same story.

Figure 2.1: School-level difference between standardized TIMSS scores and standardized NAEP scores by NAEP score in the mathematics linking sample



SOURCE: Linking sample results based on a special sample of students receiving both NAEP and TIMSS in 2000.

Figure 2.2: School-level difference between standardized TIMSS scores and standardized NAEP scores by NAEP mean in the science linking sample



SOURCE: Linking sample results based on a special sample of students receiving both NAEP and TIMSS in 2000.

If linking-sample examinees performed similarly relative to national norms on NAEP and TIMSS, these graphs would show schools roughly equally above and below the zero line on the *y*-axis. Instead, we see that virtually all schools fall below the zero line, suggesting that examinees in the linking sample perform more poorly on TIMSS relative to the national norm than they do on

NAEP relative to the national norm. We note that the same pattern holds across the range of NAEP scores, though it appears more pronounced for schools that performed better on NAEP.

Weighting

Although this sample is not a representative probability sample, it was designed to cover a range of characteristics in order to allow for an average linking function applicable to all subgroups. To make the actual sample more similar to the overall population, we have poststratified the sample by ethnicity, gender, and school type (public or private) and applied weights to bring the proportion in each group up to national averages based on the 2000 national NAEP samples.

Ethnicity used in this study is coded in three categories: black, Hispanic, and white and other. White students are combined with the rest of students with other races. School type includes two categories, public and private. The combination of three background variables results in 12 weight classes. Appendix A presents the weights used in all analyses of the linking sample.

CHAPTER III
PROJECTION LINKING

INTRODUCTION

This chapter describes our technical approach to linking the NAEP and TIMSS assessments using the data from the linking sample and presents the results of that analysis. Here, we describe

- the projection linking function;
- our approach to estimating the parameters of the linking function;
- our approach to estimating the variance of those parameter estimates, and
- the results of these analyses.

Below, we show that the linkage based on the linking sample provides reasonably precise estimates of the linking parameters. However, results in the next chapter demonstrate bias in these estimates that exceeds their putative standard errors. The next chapter also explores possible explanations.

This chapter specifies a linear relationship between the TIMSS and NAEP assessments. Having the benefit of the linking sample, we can specify the regression function $y_i = a + bx_i + u_i$, where y_i is the TIMSS score for the examinee *i*, x_i is the corresponding NAEP score, and u_i is a random error term. Linking using a regression function is a type of projection linking. For the current presentation, we assume a constant error variance across the population (say, σ_u), an assumption that we relax below. For an individual, therefore, the "linked" TIMSS score, given a score on NAEP, is actually a distribution rather than a single score. For an individual *i*, the expected TIMSS distribution has a mean of $\hat{a} + \hat{b}x_i$ and a standard deviation of σ_u ($\hat{a}$ and $\hat{b}$ denote estimates of a and b).

Two factors influence the precision of the projection of NAEP onto the TIMSS scale:

- The correlation between the NAEP and TIMSS scores[1], which is inversely related to the TIMSS variance that is not explained by NAEP (σ_u).

- The precision with which $\hat{a}$ and $\hat{b}$ are estimated.

At the individual level, σ_u is likely to be the dominant source of uncertainty. However, the impact of this factor on overall uncertainty decreases as the size of the sample increases. Both TIMSS and NAEP are designed for reporting on groups rather than individuals, so in most applications this source of error will likely prove negligible.

Several factors in addition to σ_u influence the precision with which the linking constants are estimated. These include the sample size and design, the measurement properties of each test, and the extent to which the linking model is appropriately specified. We devote much of this chapter to these issues, culminating in estimates of the full standard error around those estimates and expected confidence intervals at selected points along the scale.

Before presenting the sources of error, we discuss estimation of the linking function itself. This task is somewhat complicated by the fact that neither NAEP nor TIMSS is designed to yield reliable individual scores. Therefore, we require specialized methods to estimate the regression defining the linking function.

ESTIMATING THE PROJECTION LINKING FUNCTION

Although the linking function is a straightforward linear function, estimation is complicated by the fact that the NAEP and TIMSS instruments imperfectly measure proficiency along the NAEP and TIMSS scales. Both NAEP and TIMSS incorporate substantial measurement error at the individual level. If overlooked, this measurement error can bias linking coefficients. Appropriate analysis requires methods that explicitly model the measurement characteristics of the tests.

[1] The correlation between NAEP and TIMSS in the linking sample is .82 for mathematics and .77 for science.

Background

The measurement error in the assessments results from the fact that each student responds to a short test booklet. Both NAEP and TIMSS divide the long tests needed to cover a broad curriculum into a large number of "blocks". For both assessments only a small number of the blocks appear in each test booklet. Data from the different booklets are tied to a common scale through statistical methods based on item response theory (IRT), (Rasch, 1960; Lord, 1952; Birnbaum, 1968). Under this design, each examinee completes a single, short booklet—short enough not to exceed the student's patience for the low-stakes assessment, we hope. Across all booklets, the assessment includes many items, enough to cover the required broad curriculum.

Both NAEP and TIMSS use methods derived from marginal maximum likelihood (MML) to obtain unbiased estimates of the target statistics (average proficiencies for subgroups). Marginal maximum likelihood estimation provides estimates of marginal parameters without assigning point estimates to individuals, but rather by representing the target variable as something resembling a probability distribution. Estimates are obtained by integrating over these individual probability distributions. In a simple case, the likelihood function for individual *i* is given by

$$L_i(\mu,\sigma \mid \mathbf{z}_i) \propto \int p(\mathbf{z}_i \mid \theta) f(\theta \mid \mu,\sigma) d\theta \qquad (1)$$

where θ represents the proficiency score, μ is the population mean of the proficiency distribution, σ is the standard deviation of that distribution, and $\mathbf{z}_i$ is a vector of item responses for subject *i*. The function *p*(.) is the product over items of the probability of the observed item response, given the functional form and parameters of the IRT model. When the distribution of proficiency (θ) is normal, the mean and variance fully specify the distribution; hence, unbiased estimates of these parameters imply unbiased estimates of the entire distribution. Equation 1 explicitly models the measurement properties of the test to purge measurement error from the structural parameters.

For linking and many other applications, this model extends to a regression function by

specifying $\mu_i = \mathbf{b}'\mathbf{x_i}$, where $\mathbf{x_i}$ is a vector of individual characteristics, usually including a constant term.

For primary reporting, NAEP and TIMSS both use the "plausible values" approach to estimation, which is based on MML regression. Plausible values are imputations drawn from a large MML regression model. These imputations resemble individual test scores and have approximately the same distribution as the latent trait being measured. Aggregate statistics computed using plausible values in place of true proficiency scores are approximately unbiased. The imputation error is captured by the variance of estimates across multiple imputations (multiple plausible values) for each individual. These imputations are drawn from an MML regression model that expresses μ in Equation 1 as $\mathbf{x_i'b}$, where $\mathbf{x_i}$ is a large vector of characteristics. In NAEP and TIMSS, this vector captures most of the variance of all of the variables gathered in the background surveys. Plausible values are a special case of Rubin's multiple imputations (Rubin, 1977).

In both NAEP and TIMSS, five plausible values are drawn for each assessed student. To obtain aggregate proficiency statistics such as the mean for the population, the statistic of interest is computed with each of the five plausible values separately. The average of the five estimates is taken as the point estimate of the statistic. The variance across the five estimates is added to an estimate of the sampling variance to yield an estimate of the total variance of the estimate. Because each additional plausible value adds to the computational load, both the NAEP and TIMSS programs have judged that drawing five plausible values (as opposed to two or nine) represents an acceptable compromise between the stability of the estimates of means and variances and the computational load—both of which increase as the number of plausible values drawn increases.

Estimation

To estimate the regression coefficients for the linking model, we estimated a MML regression model with a NAEP plausible value as the independent variable and the latent TIMSS

proficiency as the dependent variable. The latent TIMSS proficiency is observed only through item responses and the IRT parameter estimates that link these responses to the underlying latent trait. Because we used NAEP plausible values as predictors in this model, we estimated the model five separate times—once using each NAEP plausible value. As mentioned above, the average of these five sets of estimates yields the point estimates, and the variance among the five estimates contributes the component of variance due to imputation error.

SOURCES OF ERROR

The projection linkage is subject to sources of error that include

- sampling error;
- measurement error;
- model specification error; and
- imputation error.

We estimated standard errors around our linking coefficients and projections in two parts. The first part used a robust Taylor-series approximation to capture the first three components of error. To capture the imputation error due to the use of NAEP plausible values as predictors, the second part repeated the estimates five times and calculated the variance among the estimates.

Sampling, measurement, and specification errors

A popular and theoretically well established method to estimate the variance of a complicated (nonlinear) statistic based on data from a complex sample survey is to use Taylor-series methods to linearize the statistic. Then, standard variance estimation methods appropriate for a linear statistic for the particular sample design are applied. This procedure is commonly called a Taylor-series approximation of the variance of a nonlinear statistic in a complex sample (Wolter, 1985).

Binder (1983) shows that the robust Taylor-series approximation to the components of error

due to sampling, measurement, and model specification may be estimated fairly simply on the basis of the score and Hessian matrices of the estimated population log likelihood, where the sample estimate of the population log likelihood is

$$LnL = \sum_{i=1}^{n} \log(L_i) w_i \,. \qquad (2)$$

Here, L_i represents an individual likelihood in the form of Equation 1, and w_i is a weight representing the inverse of the probability of selection into the sample.

Binder's method begins by defining $W_k(\theta) = \sum_i h_k(y_i, \theta) - v_k(\theta) = 0$, where k={1,2,…,M} for M parmeters in θ, $h_k(y_i, \theta)$ is a function of parameters and data, and $v_k(\theta)$ is a function of data alone. The estimate of this function from sample data, $\hat{W}(\theta)$, corresponds to the score function of Equation 2. The equation $\hat{W}(\theta) = 0$ also corresponds to what Godambe (1960) calls an estimating equation.

Binder expands $\hat{W}(\hat{\theta})$ in a Taylor series around the point $\hat{\theta} = \theta$ to obtain an approximation of the variance of the estimator that satisfies the estimating equation. Define $\hat{J}(\theta) = \frac{\partial \hat{W}(\theta)}{\partial \theta}$, and the first-order Taylor expansion is given by $\hat{W}(\hat{\theta}) = 0 \approx \hat{W}(\theta) + \hat{J}(\theta)(\hat{\theta} - \theta).$ Hence $\hat{\theta} - \theta \approx -\hat{J}^{-1}(\theta)\hat{W}(\theta).$ Substituting estimates for unknown population parameters, Binder arrives at

$$\hat{V}(\hat{\theta}) \approx [\hat{J}^{-1}(\hat{\theta})]\hat{\Sigma}(\hat{\theta})[\hat{J}^{-1}(\hat{\theta})]' \qquad (3)$$

where $\hat{\Sigma}(\hat{\theta})$ is the estimated covariance matrix of the score functions. Note that $\hat{\Sigma}(\hat{\theta})$ is simply the estimated variance/covariance matrix of a set of population totals (the summed first derivatives). In a stratified, clustered, unequally weighted sample, this can usually be approximated as the appropriately weighted estimator of the stratified, between PSU variance. Also, in our estimates, we

approximate $\hat{J}^{-1} \approx [\sum_i \mathbf{g_i}\mathbf{g_i}']^{-1}$, where $\mathbf{g_i}$ is the vector of first derivatives of the likelihood function with respect to the parameters for individual *i*.

This approximate variance estimator does not require that the observations are independent, as long as the estimate $\hat{\Sigma}(\hat{\theta})$ appropriately captures the dependence. Similarly, the estimator does not require that the model fit equally well in all cases, so cross-group differences in the linking model will be reflected in larger variances around the overall linking model.

This variance estimator captures the sampling error, as well as any model misspecification that takes the form of differential fit for various subgroups. However, Appendix B reports the results of analyses that suggest that cross-group differences in the linking function are nonsignificant. The measurement error in the TIMSS instruments is effectively captured in these standard errors because the measurement model appears explicitly in the model.

Imputation error

The imputation error in NAEP, which reflects measurement error, is captured through the standard procedures for analyzing plausible values. As mentioned above, this involves calculating the target statistic once with each plausible value and including a component reflecting variance across these estimates in the overall variance estimate. Letting θ represent a target statistic and *m* representing the number of plausible values (5 in our case), this component of variance is given by

$$Var_{pv}(\theta) = \frac{m+1}{m}\sum_{i=1}^{m}\frac{(\theta-\bar{\theta})^2}{(m-1)}.$$

Prediction error

The regression model posits that the error term has a mean of zero and a variance of σ^2. In any given sample, the mean of this random component will differ from zero. If the data were drawn from a simple random sample, the variance of the predicted mean due to this component would be

σ^2/n, where *n* represents the sample size. When calculating projections, however, NAEP does not draw data from a simple random sample. NAEP's design selects students into the sample with unequal probabilities, and selections are stratified and clustered. Therefore, the expected variance of the mean error term is somewhat larger.

We obtain a reasonable approximation of this component of variance by calculating an *effective sample size* for the samples into which we project TIMSS scores. The effective sample size is the size of a simple random sample that would yield the same precision as the actual sample. It can be calculated through the *design effect* (Kish and Frankel, 1974), which divides the appropriately estimated variance of a target variable appropriately divided by the variance calculated while ignoring stratification, clustering, and weighting. When we denote the design effect *deff*, the effective sample size is $eff = \frac{n}{deff}$. Our estimate of the variance of $\bar{u}$ is $\frac{\sigma^2}{eff}$. Finally, we obtain our estimate of the design effect from the NAEP plausible values, and assume that the same ratio holds for u. Note that this component of variance depends on the sample onto which the TIMSS scores are projected and so the effective sample size will change for different samples.

RESULTS OF PROJECTION

Tables 3.1 and 3.2 present estimates of the linking function, its overall variance, and the components of this variance based on a projection linkage using data from the linking sample. In both cases, the imputation variance is about 10 percent as large as the joint sampling, measurement, and model misspecification variance. The variance due to prediction error depends on the sample onto which TIMSS scores are projected. We see that this effect is relatively small if we imagine a sample with an effective sample size of about 1,000, where the variance due to prediction error would increase the total variance by only about 2 or 3 points.

Table 3.3 gives the effective sample sizes for the nation for the 2000 NAEP assessment and

for the 12 participating states in the 2000 NAEP and the 1999 TIMSS.

Table 3.1: Mathematics linking function and variance components using projection method in linking sample

Linking Coefficients	Point Estimate	Total Variance	Components of Variance (percent of total in parentheses)		
			Variance due to sampling, measurement, model misspecification	Variance due to imputation (NAEP measurement)	Variance due to prediction error
$A+\bar{u}$		$316.7+\frac{2387.3}{eff}$			$\frac{2387.3}{eff}$
	-56.89		289.9	26.8	
B	1.96	.0041	0.0037	0.0004	--
Cov(A,B)		-1.125	-1.0281	-0.0974	--

NOTE: *eff* is effective sample size.
SOURCE: Tabulations based on a special sample of students receiving both NAEP and TIMSS in 2000.

Table 3.2: Science linking function and variance components using projection method in linking sample

Linking Coefficients	Point Estimate	Total Variance	Components of Variance (percent of total in parentheses)		
			Variance due to sampling, measurement, model misspecification	Variance due to imputation (NAEP measurement)	Variance due to prediction error
$A+\bar{u}$		$166.0+\frac{3260.2}{eff}$			$\frac{3260.2}{eff}$
	168.94		138.5799	27.4249	
B	2.15	.0067	0.0055	0.0012	--
Cov(A,B)		-1.030	-0.8406	-0.1793	--

NOTE: *eff* is effective sample size.
SOURCE: Tabulations based on a special sample of students receiving both NAEP and TIMSS in 2000.

Table 3.3: Effective sample sizes for the 2000 NAEP samples for the nation and for the states participating in both NAEP and TIMSS

State	Mathematics			Science		
	Actual Sample Size	Design Effect	Effective Sample Size	Actual Sample Size	Design Effect	Effective Sample Size
USA	9353	4.37	2139	9443	3.57	2645
Connecticut	2454	3.71	661	2506	4.01	625
Idaho	1971	3.06	644	1973	2.48	794
Illinois	1719	3.81	451	1753	5.47	320
Indiana	1855	3.55	522	1878	5.30	354
Maryland	2401	3.09	776	2336	3.42	684
Massachusetts	2303	3.37	684	2277	5.69	400
Michigan	1975	4.13	478	2024	5.06	400
Missouri	2329	4.81	484	2320	2.74	846
North Carolina	2354	2.46	955	2342	4.30	544
Oregon	1779	3.72	478	1751	4.12	425
South Carolina	2306	3.69	625	2298	3.36	684
Texas	2317	4.26	544	2302	4.23	544

SOURCE: Tabulations based on data from the 2000 National Assessment of Educational Progress (NAEP).

In theory, it might be helpful to disentangle the sampling error, misspecification error, and measurement error; however, these sources of error are not immediately separable. Model misspecification leads to greater error variances, which in turn influences the sampling error. We have used a variance estimator that is robust to different error variances across groups, which reflects the joint effects of these forces.

Calculating the variance of the linking function is only the first challenge. The translation of these estimates into projections requires a bit of algebra. The next chapter uses these estimates to project from NAEP to TIMSS for the 12 states that participated in both studies. This provides a validation of the projection.

CHAPTER IV
VALIDATION AND INVESTIGATIONS

This chapter presents the results from a validation study. To evaluate the quality of the linkage between NAEP and TIMSS, we used the linking function to project TIMSS scores onto state NAEP datasets from 12 states that also participated in the TIMSS benchmark study. This provided a comparison between projected and actual TIMSS scores. Findings from this study suggest a bias in the linking function, which we investigated further:

- To rule out the possibility of an error in our calculations or procedure, we conducted a separate moderation linkage based on nationally reported data and projected the results of that linkage back to the linking sample and to the participating states. Results from this investigation corroborate the inference that the examinees in the linking sample performed relatively worse than expected given their NAEP scores.

- To investigate whether the observed lower performance reflected lower motivation when TIMSS was administered, we looked for evidence that examinees were not trying as hard. Specifically, we checked whether students in the linking sample were more likely to skip questions, especially constructed-response questions that require more effort. This investigation provided evidence of lower motivation in the linking sample.

- Finally, we reviewed a recent study of the similarities and differences between NAEP and TIMSS. This study identified only relatively subtle differences between the tests. Combined with the fact that TIMSS mathematics scores showed a higher correlation with TIMSS science scores than with NAEP mathematics scores, it is unlikely that such subtle differences account for observed discrepancies.

Below, we report the results of the validation study followed by the results of the three additional investigations.

COMPARISON OF PROJECTED AND ACTUAL TIMSS SCORES FOR STATES THAT PARTICIPATED IN THE BENCHMARK STUDY AND ALSO PARTICIPATED IN STATE NAEP

The U.S. TIMSS national data and data from the 12 states participating in both the state NAEP and the TIMSS Benchmark study offer an opportunity to evaluate the efficacy of the linkage. The linkage is based on a linear equation, so when we let $\hat{\bar{y}}$ represent the projected average TIMSS

score and $\bar{x}$ represent the corresponding NAEP mean, the projected average TIMSS score is given by

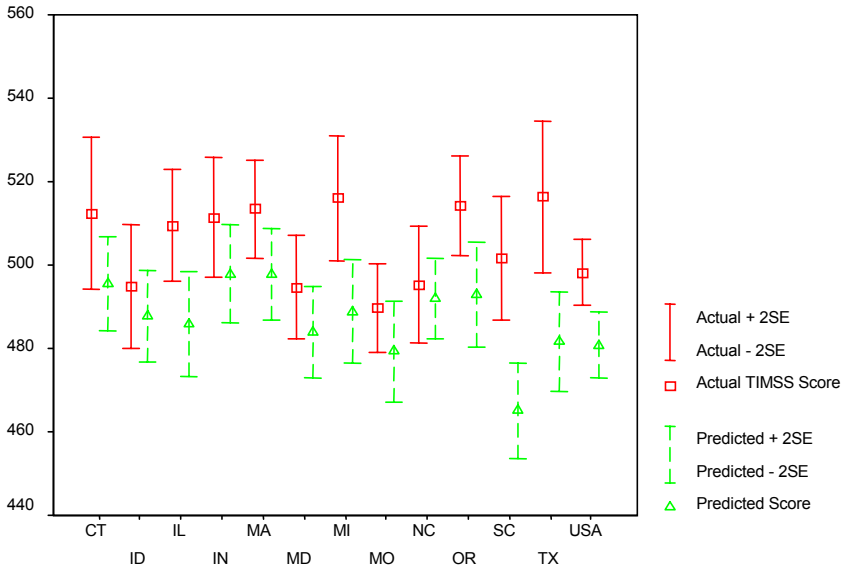
$$\hat{\bar{y}} = A + B\bar{x}.$$

We can use the variances in Tables 3.1 and 3.2 to place confidence intervals around the projections. Noting that $E(ux) = 0$, we see that the formula for the variance of the projection is given by

$$Var(y) = Var(A+u) + B^2 Var(x) + x^2 Var(B) + 2 * Cov(A,B).$$

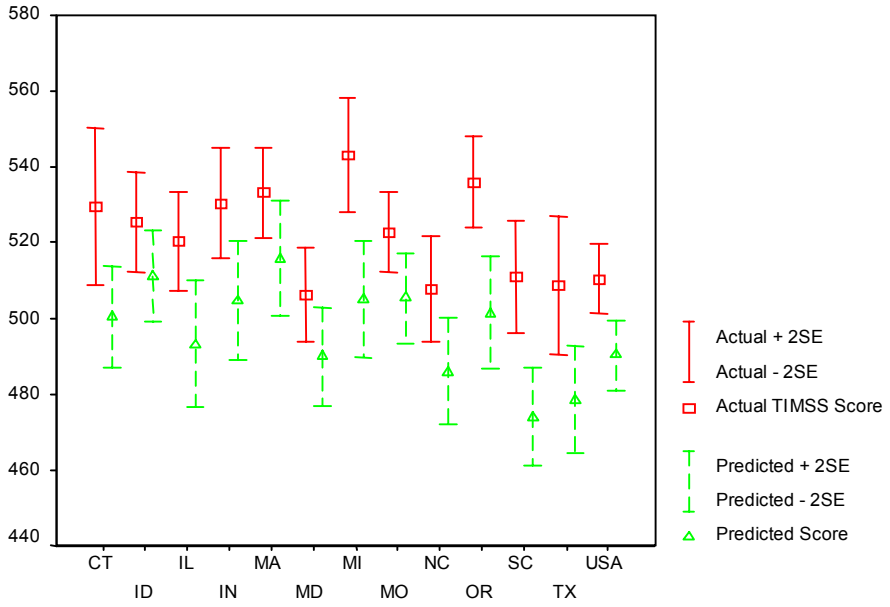
In these calculations, we use the overall variances, which incorporate sampling, measurement, imputation, model misspecification, and projection error. Figures 4.1 and 4.2 present the observed and projected TIMSS scores for the nation and for each of the 12 participating states in mathematics and science, respectively. The vertical lines mark approximate 95 percent confidence intervals around the estimates and projections.

Figure 4.1: Predicted and actual TIMSS mathematics scores with 95% standard error bands for the nation and states participating in both NAEP and TIMSS using the projection function based on the linking sample.



SOURCES: Actual values based on data from the 1999 IEA Trends in International Mathematics and Science Study (TIMSS-99). Predicted values based on applying the projection linkage to the 2000 NAEP results.

Figure 4.2: Predicted and actual TIMSS science scores with 95% standard error bands for the nation and states participating in both NAEP and TIMSS using the projection function based on the linking sample.



SOURCES: Actual values based on data from the 1999 IEA Trends in International Mathematics and Science Study (TIMSS-99). Predicted values based on applying the projection linkage to the 2000 NAEP results.

These figures make clear that this projection did not provide an acceptable linkage. The confidence bands of actual TIMSS estimates and the TIMSS projections from NAEP data rarely overlap. The projections are consistently lower than what we actually observe.

The observed bias implies a different relationship between performance on TIMSS and NAEP in the linking sample than in the operational TIMSS and NAEP studies. Despite a considerable effort to make the testing conditions as comparable as possible, some residual differences remain, which might account for the observed discrepancy:

- Operational TIMSS tests students in intact classrooms. NAEP samples students from classrooms and extracts them from the classroom for testing. Because the linking sample was a subset of the NAEP sample, the linking administration of TIMSS removed students from the classroom. This difference may have had an effect on student performance.

- Examinees responded to TIMSS about a month or so after NAEP. It is possible that actually experiencing NAEP helped them internalize the truly low stakes of the test,

which would have been apparent when they never got any feedback about performance. This realization may have reduced their motivation for the subsequent administration of TIMSS.

We might expect that either of these factors would leave other traces. In particular, nervous or less-motivated students would likely skip more items, particularly the constructed-response items that take a bit more effort to answer. We investigate this possibility below, after an additional analysis to confirm that TIMSS functioned differently in the linking sample than in the main sample.

EVALUATION OF MODERATION LINKAGE

The last NAEP-TIMSS linking study (Johnson, 1998) used statistical moderation. Statistical moderation essentially transforms one test to have the same mean and variance as the test to which it is being linked. This form of linkage is considerably weaker than a projection method because it does not rely on correlation between the assessments at all.

Here, we conduct a statistical moderation using the same approach used by Johnson (1998) to create a link between the operational national NAEP and TIMSS. Projecting this link back into the linking sample provides further evidence that the relationship between these tests is different in the operational tests than in the linking sample.

The moderation linking function can be expressed as

$$y = A + B * x \qquad (4)$$

where *y* is the transformed value on the TIMSS scale for a given value of x on the NAEP scale. The point estimates for the moderation linkage are simple: $B = \frac{\sigma_T}{\sigma_N}$ *and* $A = \mu_T - B\mu_N$, where μ_N and σ_N are the mean and standard deviation for the national NAEP data and where μ_T and σ_T are the mean and standard deviation for the national TIMSS data.

The reported means and standard deviations from the NAEP and TIMSS national public

school data are shown in Table 4.1. Table 4.2 presents the moderation linking constants based on those means and standard deviations.

Table 4.1: Means and standard deviations for national samples of US public school students, 1999 TIMSS and 2000 NAEP

Subject	TIMSS		NAEP	
	Mean	SD	Mean	SD
Mathematics	498.2	88.4	274.4	37.4
Science	510.4	98.0	149.2	36.2

SOURCES: National tabulations based on data from the 1999 IEA Trends in International Mathematics and Science Study (TIMSS-99) and the 2000 National Assessment of Educational Progress (NAEP).

Table 4.2: Estimates of moderation linking parameters based on national samples of public school students

Subject	A	B
Mathematics	-151.08	2.37
Science	106.88	2.70

SOURCES: National tabulations based on data from the 1999 IEA Trends in International Mathematics and Science Study (TIMSS-99) and the 2000 National Assessment of Educational Progress (NAEP).

Johnson (1998) uses Taylor-series linearization methods (Wolter, 1985) to derive the following approximation for the variance of an estimate based on statistical moderation, the derivation appears as Appendix C.

$$Var(\hat{\bar{y}}) \approx \hat{B}^2 Var(\bar{x}) + K_0 + K_1\bar{x} + K_2\bar{x}^2, \qquad (5)$$

where

$$K_0 = \hat{\Sigma}_{AA} = \hat{B}^2 Var(\hat{\mu}_N) + Var(\hat{\mu}_T) + \hat{\mu}_N^2 \hat{B}^2 \left\{ \frac{Var(\hat{\sigma}_T)}{\hat{\sigma}_T^2} + \frac{Var(\hat{\sigma}_N)}{\hat{\sigma}_N^2} \right\},$$

$$K_1 = 2\hat{\Sigma}_{AB} = -2\hat{\mu}_N \hat{B}^2 \left\{ \frac{Var(\hat{\sigma}_T)}{\hat{\sigma}_T^2} + \frac{Var(\hat{\sigma}_N)}{\hat{\sigma}_N^2} \right\}, \text{ and}$$

$$K_2 = \hat{\Sigma}_{BB} = \hat{B}^2 \left\{ \frac{Var(\hat{\sigma}_T)}{\hat{\sigma}_T^2} + \frac{Var(\hat{\sigma}_N)}{\hat{\sigma}_N^2} \right\}.$$

Here $\hat{\mu}_N, \hat{\sigma}_N$ and $\hat{\mu}_T, \hat{\sigma}_T$ are the estimated mean and standard deviation for NAEP and TIMSS from the national samples used in the moderation linkage; $\bar{x}$ represents the mean NAEP score in an independent sample for which TIMSS scores are desired.

The variance estimates used in these calculations are the published estimates, which capture sampling, measurement, and imputation error. Analyses similar to those reported in Johnson (1998) suggest that misspecification error was negligible.

Table 4.3 compares the estimated TIMSS score from the moderation linkage with the observed TIMSS score in the linking sample. Once more, the confidence intervals do not overlap for the linking sample, suggesting that TIMSS functions differently in the linking sample than in the operational sample.

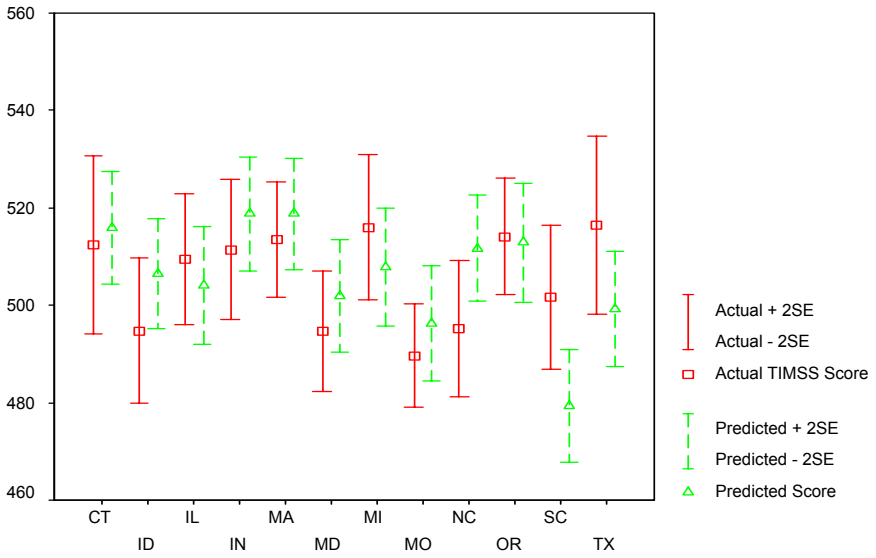
Table 4.3: Evaluation of moderation linkage as applied to the linking sample

Linking Sample	Estimated average TIMSS score	Standard error of estimated score	Observed average TIMSS score	SE of observed score	Difference between estimated and observed	Standard error of difference	z-score
Mathematics	497.5	5.0	479.9	5.0	17.6	7.1	2.5
Science	521.3	5.7	497.9	5.5	23.4	7.9	3.0

SOURCES: Moderation linkage based on national tabulations of data from the 1999 IEA Trends in International Mathematics and Science Study (TIMSS-99) and the 2000 National Assessment of Educational Progress (NAEP). Moderation linkage applied to a special linking sample of students who took both NAEP and TIMSS in 2000.

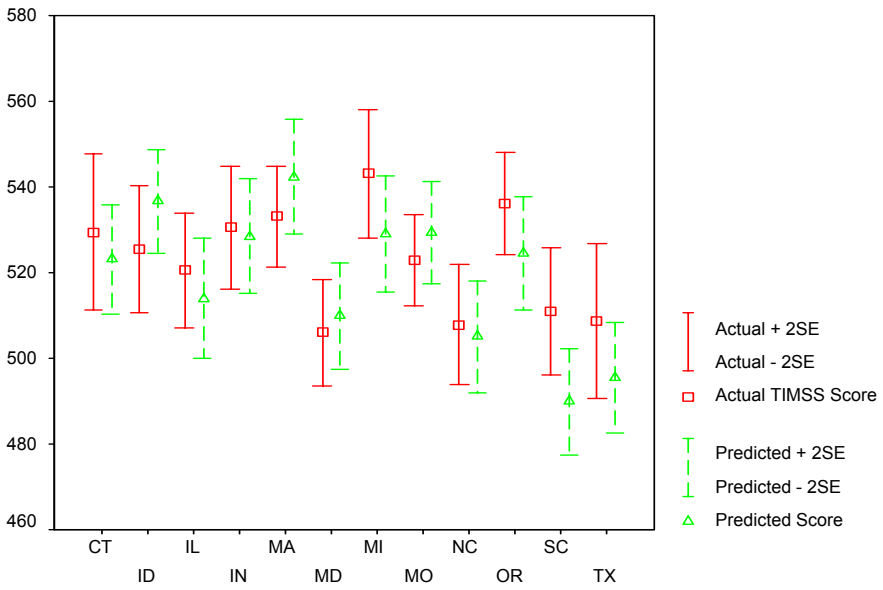
The moderation linkage does a good job of predicting TIMSS scores in the participating states. Figures 4.3 and 4.4 present the predicted and actual scores for mathematics and science (respectively) for each state, again depicting the 95 percent confidence intervals as vertical lines. Here, the confidence intervals typically overlap, suggesting that the moderation linkage does a good job predicting actual TIMSS performance.

Figure 4.3: Predicted and actual TIMSS mathematics scores with 95% standard error bands for the states participating in both NAEP and TIMSS using the moderation linking function based on the national NAEP and TIMSS data



SOURCES: Actual values based on data from the 1999 IEA Trends in International Mathematics and Science Study (TIMSS-99). Predicted values based on applying the moderation linkage to the 2000 NAEP results.

Figure 4.4: Predicted and actual TIMSS science scores with 95% standard error bands for the states participating in both NAEP and TIMSS using the moderation linking function based on the national NAEP and TIMSS data



SOURCES: Actual values based on data from the 1999 IEA Trends in International Mathematics and Science Study (TIMSS-99). Predicted values based on applying the moderation linkage to the 2000 NAEP results.

This finding presents a bit of a conundrum. The state TIMSS results based on the moderation linkage implies that if we were to give a set of students NAEP and TIMSS, the linking function would accurately (but not very precisely) predict the TIMSS performance from NAEP. This, however, was not the case. The moderation linkage proved quite inaccurate in our linking sample. The remainder of this chapter investigates possible explanations for this discrepancy.

EVALUATION OF MOTIVATIONAL DIFFERENCES

One of the plausible explanations for the drop of the TIMSS scores for the linking sample relative to the national sample is the lack of motivation on the part of the students. The linking sample consists of schools that participated in the 2000 NAEP and volunteered to participate in the linking study. It was the second low-stakes test administered within a two-month period, and the students' experience with NAEP may have confirmed the low stakes the nature of these tests. All of this may have suppressed students' motivation on the TIMSS.

We investigated this hypothesis by examining the rate at which students skipped items on the TIMSS in the linking sample and the national sample. Often, less-motivated students will simply refuse to answer more of the test items, particularly constructed-response items that require a bit more effort. Therefore, a higher proportion of omitted items would support the hypothesis that students were less motivated for TIMSS in the linking sample than the overall sample.

We calculated the omission rate as the ratio of missing responses (including items omitted and not reached) to the total items presented. We calculated the omission rates for all the items together, and separately for constructed-response items. Tables 4.4 and 4.5 present comparisons of the omission rates for the linking and national samples in mathematics and science, respectively.

Table 4.5: Comparison of omission rates between the linking sample and the national sample by type of item for mathematics TIMSS items

	Linking sample	National sample	Difference	Standard error of difference	Pr > z
All Items					
Number of items	38.8	38.0			
Number of omissions	1.0	0.8			
Omission Rate*100	2.5	2.0	0.5	0.22	0.02
Constructed response items					
Number of items	6.1	6.3			
Number of omissions	0.7	0.6			
Omission Rate*100	9.2	8.5	0.7	0.99	0.24
Multiple choice items					
Number of items	32.5	31.9			
Number of omissions	0.4	0.2			
Omission Rate*100	1.1	0.7	0.3	0.12	0.003

SOURCES: National tabulations based on data from the 1999 IEA Trends in International Mathematics and Science Study (TIMSS-99). Linking sample tabulations based on a special sample of students receiving both NAEP mathematics and TIMSS in 2000.

Table 4.6: Comparison of omission rates between the linking sample and the national sample by type of item for science TIMSS items

	Linking sample	National sample	Difference	Standard Error of difference	Pr > z
All Items					
Number of items	31.2	30.7			
Number of omissions	0.7	0.5			
Omission Rate*100	2.2	1.7	0.6	0.26	0.02
Constructed response items					
Number of items	5.2	5.1			
Number of omissions	0.5	0.4			
Omission Rate*100	7.6	6.5	1.0	1.00	0.15
Multiple choice items					
Number of items	25.9	25.5			
Number of omissions	0.3	0.2			
Omission Rate*100	1.0	0.6	0.4	0.17	0.01

SOURCES: National tabulations based on data from the 1999 IEA Trends in International Mathematics and Science Study (TIMSS-99). Linking sample tabulations based on a special sample of students receiving both NAEP science and TIMSS in 2000.

These tables show that the students in the linking sample skipped or failed to reach significantly more items than their counterparts in the national study. In both subjects, the overall omission rates and the omission rates are significantly higher in the linking sample. Although the higher omission rates for constructed-response items are not statistically different from the national figures, these too tend to be higher in the linking sample. The small numbers of constructed-response items make the omission rates on these items much less stable. On the whole, these data suggest greater omission in the linking sample than in the national sample.

These omission rates alone are not sufficient to account for the discrepancy in performance; however, they point to a potential lagging motivation or reduced concentration among examinees in the linking sample when they were taking TIMSS.

This analysis suggests that something either distracted or demotivated the linking sample students when they were taking TIMSS. Above, we offered two possible causes: 1) their recent experience with NAEP could have convinced them of the absence of consequences associated with the test; or 2) removal from their classroom may have distracted or distressed them, impeding their concentration on the test. In either case, the analysis suggests that TIMSS did, in fact, function differently in the linking sample than it did in the national sample. The magnitude of the differences in performance should serve as a warning about the fragility of assessment results to effects of seemingly irrelevant factors.

EVALUATION OF CONTENT VS. CONTEXT THROUGH CORRELATION

An unintuitive phenomenon has been observed in the correlations among the test scores in the linking sample; namely, the TIMSS and NAEP mathematics are correlated at .82, whereas the TIMSS mathematics and science correlated at .87. One possible explanation is that the contents of TIMSS and NAEP mathematics were different; that they essentially tested different mathematics proficiency, which led, therefore, to the lower correlation.

However, this explanation contradicts the findings by a study comparing the contents of the three assessments: NAEP 2000, TIMSS-99, and PISA (Programme for International Student Assessment) (Nohara, 2001). The study asked three questions:

1. Do the assessments cover the same topics?
2. Do the assessments ask the same type of questions?
3. Do the assessments ask the students to use similar types of thinking skills?

The study concluded that NAEP and TIMSS-99 are very similar in content coverage. For example, both assessments focused heavily on number sense, properties, and operations. Thirty-two percent of NAEP items addressed this topic, compared with 46 percent of the TIMSS-99. As for the

item types, 60 percent of the NAEP items were multiple choice and 16 percent were short-answer. For TIMSS-99, 77 percent were of multiple choice items, and 20 percent were short answer items. The NAEP assessment had the higher percentage of extended constructed-response items and was considered more difficult.

TIMSS-99 had a higher percentage of items requiring computation relative to NAEP. NAEP, however, had a higher percentage of items that required multistep reasoning and interpretation of figures and charts." However, the overall percentages were similar for the two assessments. The percentage of "interpretation of figures" items was the highest for both assessments. In all, the study concluded that NAEP and TIMSS had the same content coverage, with NAEP the more balanced of the two. Both assessments included the same types of questions, although NAEP had more extended constructed-response items than TIMSS-99. Finally, both assessments asked students to use similar kinds of thinking skills. The study also concluded that NAEP is slightly more difficult than the TIMSS-99.

The findings of Nohara's study indicate that we cannot attribute the lower correlation between the TIMSS and NAEP mathematics to content difference. The other plausible explanation then is the difference in context. First, we can generally observe that mathematics performance and science performance correlate highly. In addition, the TIMSS mathematics and science assessments were administered at the same time in the same environment (i.e., the same context), whereas the NAEP was administered at different times and in different manners from TIMSS. The administrations of the NAEP and TIMSS differed in two important ways. For NAEP, this was the first time that students encountered the large-scale low-stake assessment. In addition, the students were not in their own classrooms. The TIMSS was the students' second encounter in less than two months for the linking sample. This memory was fresh and the situation was more familiar. We speculate that differences in context outplayed similarities in content to contribute to the lower correlation between TIMSS and NAEP mathematics relative to TIMSS mathematics and science.

CHAPTER V
DISCUSSION

SUMMARY

This study attempted to link the TIMSS mathematics and science assessments to the NAEP assessments in the same subjects. The primary linkage used a projection method, which drew data from a sample of students to whom both assessments were administered. We attempted to validate the linkage by projecting TIMSS scores from NAEP scores in a dozen states that participated in both studies, only to find that the projections were substantially off the mark.

We developed a secondary linkage based on nationally reported numbers, using a statistical moderation approach. This approach provides a fairly weak linkage, since it essentially only involved putting two different tests on the same scale so that the mean and standard deviation matched. Such a linkage could be conducted between entirely unrelated tests, since the method uses no information about the correlation between the tests being linked. This moderation linkage did a decent job of projecting TIMSS scores from NAEP scores in the 12 states that participated in both studies, but failed to predict the TIMSS score in the linking sample.

The analyses in the report show that the TIMSS assessments functioned differently in the linking sample than they did in the national and state samples. We identified two contextual differences that may contribute to this discrepant performance:

- The national TIMSS was conducted in intact classrooms, whereas the linking study followed the NAEP approach of sampling students school wide and gathering them together in a room for testing. We posit that this may unnerve students or highlight for them the fact that the assessment results will not affect them personally.
- TIMSS was administered to the linking sample several months after NAEP, which gave the examinees firsthand experience with the lack of consequences for NAEP. This personal understanding of the fact may have relieved the students of any motivation that they may have otherwise had.

Analysis found that the rate at which students failed to answer TIMSS items (the omission rate) was higher in the linking sample than in the national sample, supporting the idea that students were less motivated or concentrating less well in the linking study. Although our analysis supported this hypothesis, it remains possible that other, as yet unidentified, forces were at work.

EFFECTS OF CONTEXT

The main implication of this linking study is the sensitivity of assessment results to context. The linking study was carefully designed to match the administration contexts as closely as possible. Even the administration window of the TIMSS was designed to match the administration window of the national TIMSS. However, if our speculation is correct, other subtle contextual differences may have had enough influence to depress scores by about 20 TIMSS points.

The pattern of correlations among test scores in the linking sample underscores the impact of context. The correlation between TIMSS and NAEP mathematics performance is 82 and the correlation between TIMSS mathematics and TIMSS science is .87. Thus, two tests with very disjoint content but administered in the same context correlate more strongly than two tests with very similar content, but administered in subtly different context. The difference in the context between NAEP and TIMSS in the linking sample was limited to the administration window and the facts that examinees had already experienced NAEP when they took TIMSS and that the TIMSS assessment took place in non-intact classrooms.

To the extent that impact of context rivals that of content, the issues of generalizability and validity converge. Statements about test performance rarely include caveats such as “if assessed before taking any other large scale assessment…” or “when tested in their own classrooms…” If these sorts of contextual factors are as important as the current results suggest, validity will require more generalizable approaches to assessment. Alternatively, statements about proficiency could be made valid by confining them to contexts so narrow that they become uninteresting.

Although issues relevant to the state of the discipline ought to not be lost, they are also unlikely to be resolved immediately. Hence, we turn our attention to the more immediate concern of how we might design a linking study to better effect.

DESIGNING THE NEXT LINKING STUDY

The current linking study was designed so that the same students responded to both NAEP and TIMSS using the same instruments as in the national NAEP and TIMSS. Additionally, to the extent possible, the same administration procedures were followed in the administration of TIMSS to the linking sample as were used in TIMSS-99 (since the students in the linking sample were selected from those taking NAEP, the NAEP procedures were automatically identical). Nevertheless, the projection linkage using the linking sample data failed.

There were two differences between the administration conditions experienced by the students in the linking sample when they took TIMSS and those for students in the 1999 TIMSS. Both changes were required by the linkage study design in which a subsample of the students selected for NAEP were then administered TIMSS. This design led to non-intact classroom testing in the linking sample (as opposed to intact classroom testing in TIMSS-99). It also led to potential training on the low-stakes nature of TIMSS from having experienced no consequences from the NAEP testing some months earlier.

Either or both of these factors might have contributed to the observed failure of the linking functions—lower motivation due to experience with NAEP, and lower performance due to non-intact-classroom testing. It is possible that these two factors work together, and it is also possible that neither has much bearing on the issue. In any event, we ought to discern the problems before the next attempt to link two large-scale assessments.

Studies could be designed to evaluate the impact of each of the posited factors. To evaluate the effect of intact classrooms, randomly equivalent groups could take the test under each condition.

The working hypothesis holds that the intact classroom group would perform better.

To evaluate the impact of a prior low-stakes assessment, the spiral of NAEP booklets could be altered in a sample of schools to withhold a set of blocks from those schools. The withheld blocks could be administered in the selected school under standard NAEP conditions a couple of months later. We hypothesize that we would see lower performance, despite the intervening learning.

The design of the next linking study would depend on the outcomes of these two evaluation studies:

- If only "intact classrooms" had an effect, we could use a design similar to the one reported here, but test all students in intact classrooms when TIMSS is administered for linking. Only responses from NAEP respondents would have to be scored and analyzed for the linkage.

- If only "prior testing" had an effect, TIMSS blocks could be spiraled with operational NAEP in a subsample of schools, effectively administering both tests simultaneously.

- If both effects exist, TIMSS blocks could be spiraled with operational blocks, and an adjustment factor for non-intact-classrooms could be applied to the linkage. The adjustment factor (along with its standard error) could come from the evaluation study. Foresight suggests that the sample for that study be large enough to yield sufficiently precise estimates of the "intact classroom effect." Too small a sample would risk standard errors around the adjustment that could become the dominant term in the linkage function.

- Finally, if neither effect exists, we ought to return to the data and other research to identify and test other possible sources before proceeding with another linking study.

The evaluation studies proposed here have implications beyond the next NAEP-TIMSS linkage. Understanding the contextual factors that influence student performance on large-scale assessments can ultimately lead to improved designs for the assessments themselves.

APPENDIX A
WEIGHTS FOR LINKING SAMPLES

School Type	Gender	Race	Proportion in National Sample	Proportion in Linking Sample	Weight
Mathematics					
Public	Male	White	0.329	0.179	1.838
		Black	0.059	0.057	1.017
		Hispanic	0.066	0.043	1.535
	Female	White	0.314	0.195	1.610
		Black	0.065	0.063	1.038
		Hispanic	0.065	0.030	2.150
Private	Male	White	0.043	0.168	0.256
		Black	0.004	0.021	0.204
		Hispanic	0.005	0.028	0.173
	Female	White	0.042	0.182	0.231
		Black	0.004	0.016	0.250
		Hispanic	0.004	0.018	0.222
Total			1.000	1.000	1.000
Science					
Public	Male	White	0.329	0.176	1.869
		Black	0.062	0.051	1.213
		Hispanic	0.064	0.038	1.679
	Female	White	0.314	0.178	1.760
		Black	0.066	0.057	1.150
		Hispanic	0.066	0.036	1.808
Private	Male	White	0.044	0.194	0.226
		Black	0.003	0.014	0.220
		Hispanic	0.004	0.026	0.173
	Female	White	0.041	0.193	0.215
		Black	0.003	0.010	0.297
		Hispanic	0.004	0.026	0.166
Total			1.000	1.000	1.000

NOTE: These weights, when applied to the linking sample, are designed to make the relative proportions within the 12 cells in the linking sample match those in national NAEP.
SOURCES: Computations based on national NAEP 2000 data and data from the sample of students taking both NAEP and TIMSS.

APPENDIX B
CROSS-GROUP DIFFERENCES IN THE PROJECTION LINKING FUNCTION

To be useful, the link between NAEP and TIMSS should be the same for various subpopulations. That is, the function linking the two assessments should be the same for boys as it is for girls and for members of various ethnic categories. To the extent that the link is consistent across the subpopulations, there is increased confidence in the goodness of the link.

To verify the stability of the projection-based link across subpopulations, we partitioned the linking sample into 6 mutually exclusive categories defined by cross gender with ethnicity (black, Hispanic, white+other). The projection linking function was then fit separately for each of the six subpopulations and the variance of the slopes and of the intercepts was computed. These cross-group variances form the measures of model misspecification, one for the slopes and one for the intercepts, and are denoted $\hat{V}ar(Misspec)$ in Table C.1 below. Also shown in the table are measures of variability of the estimates, due both to sampling error and to measurement error. These variability measures are denoted $\hat{V}ar(S+M)$ in the table. The ratio of the first to the second gives an indication of how large the effect of model misspecification is relative to the error due to sampling and measurement error. The ratio is approximately distributed like an F distribution. Since the ratio is less than 1 in all cases, it can be seen that, for both mathematics and science, and for both the intercepts and the slope, the misspecification error is nonsignificant.

Table B.1: Evaluation of cross-group differences in the projection linking function

Linking Sample	Intercept			Slope		
	$\hat{V}ar(Misspec)$	$\hat{V}ar(S+M)$	*Ratio*	$\hat{V}ar(Misspec)$	$\hat{V}ar(S+M)$	*Ratio*
Mathematics	1106.04	2001.02	0.55	1.382E-05	2.041E-05	0.67
Science	3262.21	4147.30	0.79	3.057E-05	4.180E-05	0.73

SOURCE: Results from computations using data from the linking samples of students who took both NAEP and TIMSS in 2000.

APPENDIX C
VARIANCE ESTIMATION FOR THE MODERATION LINKAGE VIA TAYLOR SERIES LINEARIZATION

The moderation linkage function can be expressed as

$$f_T = y = A + B * x$$

where *y* are the transformed values on the TIMSS scale and *x* are the values on the NAEP scale. The point estimates for the moderation linkage are simple:

$$B = \frac{\sigma_T}{\sigma_n} \quad and \quad A = \mu_T - B\mu_N.$$

Since f_T is a nonlinear function of the various means and standards deviations, a precise derivation of the variance of $\hat{y}$ is not practical. However, since both the NAEP and TIMMS samples are large, Taylor series linearization (Wolter, 1985) provides a convenient large sample approximation to the variance:

$$\left[\frac{\partial f_T}{\partial x} \; \frac{\partial f_T}{\partial \hat{A}} \; \frac{\partial f_T}{\partial \hat{B}}\right] \Sigma \left[\frac{\partial f_T}{\partial x} \; \frac{\partial f_T}{\partial \hat{A}} \; \frac{\partial f_T}{\partial \hat{B}}\right]^T,$$

where the partial derivatives are evaluated at x, $\hat{A}$ and $\hat{B}$ respectively, the superscript T denotes matrix transpose, and Σ is the matrix

$$Var([x\ \hat{A}\ \hat{B}]) = \begin{bmatrix} \Sigma_{xx} & 0 & 0 \\ 0 & \Sigma_{AA} & \Sigma_{AB} \\ 0 & \Sigma_{AB} & \Sigma_{BB} \end{bmatrix},$$

where $\Sigma_{xx} = Var(x), \Sigma_{AA} = Var(\hat{A}), \Sigma_{BB} = Var(\hat{B}), \Sigma_{AB} = Cov(\hat{A}, \hat{B}),$ and where the covariances between x and $\hat{A}$ and x and $\hat{B}$ are both zero since x is from a sample independent of those used to construct the estimates of $\hat{A}$ and $\hat{B}$.

Since

$$\left[\frac{\partial f_T}{\partial x}\ \frac{\partial f_T}{\partial \hat{A}}\ \frac{\partial f_T}{\partial \hat{B}}\right]=[\hat{B}\,1\,x],$$

one has

$$Var(\hat{y}) \approx \hat{B}^2\Sigma_{xx} + \Sigma_{AA} + 2x\Sigma_{AB} + x^2\Sigma_{BB.}$$

Estimates of Σ_{AA}, Σ_{AB}, and Σ_{BB} can be obtained by expressing $\hat{A}$ and $\hat{B}$ in terms of $\hat{\mu}_T$, $\hat{\sigma}_T$, $\hat{\mu}_N$, and $\hat{\sigma}_N$ and applying the delta method to the result.

Let $\Xi=[\hat{\mu}_T\ \hat{\sigma}_T\ \hat{\mu}_N\ \hat{\sigma}_N]$ and $\Sigma_{\Xi\Xi}=Var([\hat{\mu}_T\ \hat{\sigma}_T\ \hat{\mu}_N\ \hat{\sigma}_N])$. Since the mean and standard deviation from the NAEP sample is independent of those from the TIMSS sample, and since a sample mean and a sample standard deviation are independent assuming normality, $\Sigma_{\Xi\Xi}$ can be conveniently and credibly taken as a diagonal matrix with diagonal elements $\{Var(\hat{\mu}_T), Var(\hat{\sigma}_T), Var(\hat{\mu}_N), Var(\hat{\sigma}_N)\}$. As

$$\begin{bmatrix}\Sigma_{AA} & \Sigma_{AB}\\ \Sigma_{AB} & \Sigma_{BB}\end{bmatrix} \approx \begin{bmatrix}\frac{\partial \hat{A}}{\partial \Xi}\\ \frac{\partial \hat{B}}{\partial \Xi}\end{bmatrix} \Sigma_{\Xi\Xi} \begin{bmatrix}\frac{\partial \hat{A}}{\partial \Xi}\\ \frac{\partial \hat{B}}{\partial \Xi}\end{bmatrix}^T$$

and

$$\frac{\partial \hat{A}}{\partial \Xi} = \left[1 \quad -\frac{\hat{\mu}_N}{\hat{\sigma}_N} \quad -\frac{\hat{\sigma}_T}{\hat{\sigma}_N} \quad \frac{\hat{\mu}_N}{\hat{\sigma}_N}\frac{\hat{\sigma}_T}{\hat{\sigma}_N}\right]$$

$$\frac{\partial \hat{B}}{\partial \Xi} = \left[0 \quad \frac{1}{\hat{\sigma}_N} \quad 0 \quad -\frac{1}{\hat{\sigma}_N}\frac{\hat{\sigma}_T}{\hat{\sigma}_N}\right],$$

some algebra produces

$$Var(\hat{y}) \approx \hat{B}^2 \{Var(x) + Var(\hat{\mu}_N)\} + Var(\hat{\mu}_T)$$

$$+ (x - \hat{\mu}_N)^2 \hat{B}^2 \left\{ \frac{Var(\hat{\sigma}_T)}{\hat{\sigma}_T^2} - \frac{Var(\hat{\sigma}_N)}{\hat{\sigma}_N^2} \right\}.$$

Since $Var(\hat{y})$ depends on x and $Var(x)$, it is convenient to reexpress $Var(\hat{y})$ as

$$Var(\hat{y}) \approx \hat{B}^2 Var(x) + K_0 + K_1 x + K_2 x^2,$$

where

$$K_0 = \hat{\Sigma}_{AA} = \hat{B}^2 Var(\hat{\mu}_N) + Var(\hat{\mu}_T) + \hat{\mu}_N^2 \hat{B}^2 \left\{ \frac{Var(\hat{\sigma}_T)}{\hat{\sigma}_T^2} + \frac{Var(\hat{\sigma}_N)}{\hat{\sigma}_N^2} \right\},$$

$$K_1 = 2\hat{\Sigma}_{AB} = -2\hat{\mu}_N \hat{B}^2 \left\{ \frac{Var(\hat{\sigma}_T)}{\hat{\sigma}_T^2} + \frac{Var(\hat{\sigma}_N)}{\hat{\sigma}_N^2} \right\}, \text{ and}$$

$$K_2 = \hat{\Sigma}_{BB} = \hat{B}^2 \left\{ \frac{Var(\hat{\sigma}_T)}{\hat{\sigma}_T^2} + \frac{Var(\hat{\sigma}_N)}{\hat{\sigma}_N^2} \right\}.$$

REFERENCES

Binder, D.A. (1983) On the variances of asymptotically normal estimators from complex surveys, *International Statistical Review, 51, 279-292.*

Birnbaum, A (1968) *Some latent trait models and their use in inferring an examinee's ability.* In Lord, F.M. and Novick, M.R. Statistical Theories of Mental Test Scores, Reading MA: Addison-Wesley.

Cohen, J. & Jiang, T. (2001) *Direct Estimation of Latent Distributions for Large-Scale Assessments with Application to the National Assessment of Educational Progress (NAEP).* Washington, DC: American Institutes for Research.

Godambe, V.P.(1960) An optimum property of regular maximum likelihood estimation. *Ann. Math. Stat*., 31, 1208-1211.

Johnson, E. G. (1998) *Linking the National Assessment of Educational Progress and the Third International Mathematics and Science Study: A Technical Report*. Washington, DC: U.S. Department of Education, National Center for Education Statistics, NCES 98-499.

Kish, L. and Frankel, M.R. (1974) Inference from complex samples, *Journal of the Royal Statistical Society, Series B*, 36, 1-22.

Lord, F. M. (1952). *A theory of test scores. Psychometric Monograph, No. 7.* Chicago: University of Chicago Press.

Nohara, D. (2001) *A comparison of the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study Repeat (TIMSS-R), and the Programme for International Student Assessment (PISA).* Washington, DC: U.S. Department of Education, National Center for Education Statistics, NCES 2001/07.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: The Danish Institute of Educational Research.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley & Sons.

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