This is to certify that the

dissertation entitled

LINKING CLASSROOM ASSESSMENT PRACTICES
TO LARGE-SCALE TEST PERFORMANCE

presented by

MICHAEL CLIFFORD RODRIGUEZ

has been accepted towards fulfillment
of the requirements for

PhD degree in MEASUREMENT
AND QUANTITATIVE METHODS

S. E. Phillips
Major professor

Date 6-30-99

LINKING CLASSROOM ASSESSMENT PRACTICES TO LARGE-SCALE TEST PERFORMANCE

By

Michael Clifford Rodriguez

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Counseling, Educational Psychology, and Special Education
Measurement and Quantitative Methods Program

1999

ABSTRACT

LINKING CLASSROOM ASSESSMENT PRACTICES
TO LARGE-SCALE TEST PERFORMANCE

By

Michael Clifford Rodriguez

What to measure and how to measure it are enduring issues in educational measurement (Ebel, 1982; Lindquist, 1936), both in terms of large-scale assessment and classroom assessment. Recent attention on accountability systems and policies within states have brought greater attention to the measurement of student outcomes and have also prompted national professional organizations (e.g., National Council of Teachers of Mathematics and American Association for the Advancement of Science) to adopt standards for the practice of assessment. Although most of the efforts to develop benchmarks and curriculum frameworks (what to measure) have a strong research base, most assessment practices promoted by these organizations (how to measure) appear to be based on anecdotal experiences. Some organizations more than others have adopted recommendations of measurement specialists (e.g., AFT, NCME, & NEA, 1992).

This project was an attempt to evaluate a larger classroom assessment system, including the role of student self-efficacy and effort in mediating the relationships between assessment practices and achievement. In part, this was based on a framework proposed by Brookhart (1997). The United States portion of the Third International Math and Science Study (TIMSS) database was used to estimate these relationships. Based on background questionnaires and achievement data from 6963 students and their mathematics teachers (including 326 teachers), a hierarchical linear model was fit to the

data. Nearly 54 percent of the variance in student mathematics scores was between classrooms while 46 percent was within classrooms.

The full HLM model accounted for 65 percent of the variance between classrooms and an additional 8 percent of the variance within classrooms. By including a composite indicator for the relative prior math achievement of students within classrooms given content of current courses, 28 percent of the variance in classroom performance was accounted for. This indicator served a combined role in accounting for the level of mathematics content covered in each class and the prerequisite skill level of students (i.e., loosely speaking, ability of students).

At the student level, mothers' education, mothers' expectations, self-efficacy, and effort had significant positive relationships to student performance, while level of uncontrollable attributions had a negative relationship to performance. At the classroom level, teachers' use of teacher-made objective tests, and their use of assessment information for grading and evaluation rather than feedback and discussion had significantly negative relationships to classroom performance.

In addition, frequent use of teacher-made objective tests at the classroom-level neutralized the positive relationship between self-efficacy and performance at the student-level while frequent use of teacher-made objective tests increased the negative relationship between uncontrollable attributions and performance. These cross-level interactions suggested that classroom assessment practices might uniquely interact with student characteristics in their role of motivating student effort and performance. A framework for classroom assessment research was also presented.

ACKNOWLEDGEMENTS

The faculty in my department have been supportive beyond the call and my acknowledgement of them here will be insignificant compared to the impact they have had on me as a person and on my training as an educational researcher. I owe great thanks to Betsy Becker, Ken Frank, Richard Houang, Irv Lehmann, David Pearson, and Mark Reckase. Tom Haladyna (Arizona State University) solidified my faith in the measurement community, as one sincerely interested in the development of thoughtful measurement specialists and the improvement of the practice of educational measurement--I look forward to our continued collaboration. I also must acknowledge the late Robert Ebel, who supported my first three years of enrollment through his generous endowed scholarship and the National Council on Measurement in Education for their financial support of my last two years.

My dissertation committee deserves a great deal of credit for allowing me to pursue an extremely complex arena for study, the middle school mathematics classroom. With their advice and encouragement, I have mapped out the beginnings of what could become a lifetime research program, of which this dissertation is a small piece. William Mehrens, Robert Floden, Teresa Tatto, and my advisor, Susan Phillips, have been outstanding sources of intellectual guidance and have raised my own expectations for performance. I particularly owe a great deal to Dr. Phillips; she has unfailingly encouraged me to pursue academic and professional activities that will carry me through academia successfully.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF TABLES IN APPENDIX A

Questions from the teacher background questionnaire:

Questions from the student background questionnaire:

LISTOF FIGURES

CHAPTER I

Problem Statement

Assessment impacts students through the practices employed by their teachers. Teachers review results of standardized tests, create tests of their own using various formats, evaluate completed student projects they developed or obtained from resource guides or textbooks, and assign work to be done outside of school. They ask questions, listen, watch, interview students, pose questions for solution by individuals or groups of students. Then, to one extent or another, teachers communicate their findings and evaluations to students, and in doing so, impact the learning process, which fully includes participation in instructional activities, self-selected learning activities, assessment activities, and subsequent feedback from teachers. Directly, assessments impact students by communicating learning goals, including the subject-matter content and thinking processes valued by their teachers.

Assessment impacts students by shaping study behaviors, and general and academic self-concepts and self-efficacy; enabling self-adjustment, enhancing academic motivation, and organizing and securing the storage of knowledge and skills. Assessment at the classroom level is clearly important. (Most of these ideas were based on research that was reviewed below.) This means that teachers must know something about assessment.

National educational organizations have been developing and promoting standards for assessment, both at the classroom level and regarding national assessments. States have adopted statewide tests in part to reform instructional and assessment

practices of their teachers. In this storm of assessment and testing standards, researchers have been trying to describe the relationship between assessment and learning or achievement.

The primary intent of this work was to evaluate the relationship between classroom level assessment practices and student performance on a large-scale assessment. It was expected that some assessment practices employed by teachers help some students and not other students. Some assessment practices help some students obtain certain outcomes and not others. For example, some classroom assessment tools may help students prepare for large-scale or standardized tests, while others help students prepare for success in college or to get certain jobs. Relevant questions become: Which practices, which outcomes, which students?

If these suppositions about the impact of assessment practices were true, strong evidence could be gathered to argue for assessment reform and the establishment of assessment standards. Policy makers would have solid ground to influence what teachers know about assessment, what they do in practice, how they are trained, and what constitutes appropriate certification and professional development activities. This should also include the improvement of assessment competence of school administrators who, as Trevisan (1999) argued, are in critical roles to support teachers and their classroom assessment activities and to help build connections between classroom assessment practices and district or state assessment activities. Trevisan found serious lack of attention by nearly every state to administrator assessment competence advocated by several national professional organizations.

Measurement specialists have continually suggested improvements in classroom measurement-related professional development. Cross and Frary (1999) recommended recently that measurement specialists attempt to communicate with a broader audience concerning the merits of best practice, particularly outside of the measurement journals (this was in regard to the prevalence of "hodgepodge" grading practices of teachers). They cited several negative consequences of limited measurement knowledge in the practice of classroom assessment.

Communication regarding the merits of best practice must be improved at all levels, including during teacher preparation and professional development, policy analysis and design, program implementation, evaluation and design of standards of practice, and evaluation of student and teacher performance. This is predicated on the value of the information communicated. For most of the questions posed earlier, little to no information exists. Requirements for certification, topics of professional development, and standards of practice are not substantially informed by evidence. This is due, in large part, to the absence of evidence regarding the impact of classroom assessment practices. With the current focus in education policy on accountability and the broad implementation of standards of practice, the need for evidence to support these efforts is at a critical high. The search for evidence to support classroom assessment reform is sparse and has not been equal to the complexity of the task. This project was developed as an effort to broaden the scope of coverage in understanding key relationships in the classroom assessment environment.

General Motivation

Nearly one year ago, the editors of *Education Week* (Edwards, 1998) posed a question to policy makers, educators, and the American public: If one school can succeed under the worst conditions, with the neediest children, how can others be permitted to fail? The second edition of their special report, *Quality Counts '98*, focused on urban education because that was where, according to the editors, the greatest gap between a state's expectations for student achievement and the reality of student achievement existed. They reviewed barriers to success and argued "the problems confronting urban school districts are bigger, costlier, more numerous, and tougher to overcome than those facing most rural and suburban systems" (Olson & Jerald, 1998a, p.9).

Michigan's coverage in *Quality Counts '98* was not as a standout performer, but as a state with large urban-nonurban school district achievement gaps. Based on results of the 1996 National Assessment of Educational Progress (NAEP) mathematics test for Michigan eighth graders, 37% of urban districts scored at basic level or higher while 74% of nonurban districts scored at basic level or higher; Michigan had the third largest urban-nonurban gap in the nation. For the 1996 NAEP science results for Michigan eighth graders, 33% of urban districts scored at basic level or higher while 72% of nonurban districts scored at this level; the fifth largest gap in the nation.

Editors of *Quality Counts '98* chose to focus on the concentration of poverty as the largest barrier to achievement. They recognized, however, that poverty was not the sole reason for the "gap" in performance. "Somehow, simply *being* in an urban school seems to drag performance down" (Olson & Jerald, 1998b, p. 10). They continued presenting statistics on poverty, high school drop-out rates, teacher qualifications and

turnover, parent involvement, violence, school size (which is confounded with district population), absenteeism, and other information regarding access to resources, politics, and governance. They also presented twelve ideas that educators and policy makers have argued are necessary for progress (summarized in Table 1).

Although they did not address assessment directly, each provides some support for the role of assessment in communicating learning goals (#1 from table), the importance of teacher competence in assessment (#4), the importance of professional development (#5), the need for school administrators and other leaders to be able to support assessment activities of teachers (#6), and communicating results to parents (#8).

Table 1
Twelve Ways to Improve Achievement and Educational Progress

Contributing Steps to Progress
1. setting clear, high expectations for all students
2. devising an accountability system based on good information
3. creating clear lines of authority; give schools freedom in exchange for accountability
4. recruiting, hiring, and retaining teachers who can enable students to reach high standards
5. building capacity at the school level to improve teaching and learning with a strong focus on better curriculum and instruction
6. creating strong leaders at the school and district levels
7. getting students the extra time and attention they need to succeed
8. improving the relationship of parents and communities with schools and educators
9. thinking small--size isn't everything, especially in big-city schools
10. providing safe and adequate school buildings
11. breaking up the monopoly on district-run schools
12. closing or reconstituting bad schools

Overall, Michigan received an A- in 1997 and a B+ in 1998 for Standards and Assessments from the editors of *Quality Counts '98*, based on the state's high standards for all children and assessments aligned with those standards. This placed Michigan thirteenth in the nation. Michigan also ranked tenth for quality of teaching, while only receiving a C- for having teachers who have the knowledge and skills to teach to higher standards.

Soon after the *Quality Counts '98* issue was released, the Detroit *Free Press* published a series of articles entitled *Testing MEAP Scores*. The series investigated the relationships between the Michigan Educational Assessment Program tests in math, science, reading, and writing, and school district demographics. A regression analysis was completed using the average percent of students in a district that achieved passing scores on the reading and math tests. The explanatory variables included demographic indicators such as percent of households where no one was a high school graduate, local unemployment rate, percent of single-parent households, and school funds per pupil, percent of students who spoke English as a second language. The primary criticism leveled against the MEAP was that it resulted in scores that were used for district versus district comparisons that were unfair. For instance, because these demographics were associated with 62% of the variance in the MEAP index as described above for urban school districts, "MEAP results show more about who's taking the tests than how well they're being taught" (Van Moorlehem, 1998a).

The *Free Press* team then calculated predicted scores computed from the regression equations for three levels of schools based on size with the corresponding R^2 (variance explained): large districts ($R^2 = 0.62$), medium sized districts ($R^2 = 0.33$), and

small sized districts ($R^2 = 0.16$). Based on differences between the predicted scores and the actual observed scores, the *Free Press* staff writers then organized schools by those that scored above what could be expected, about what could be expected, and below what could be expected. Considering MEAP scores in light of factors that were not within the control of schools, the *Free Press* "produced some surprising results."

> Detroit schools, for example, are overcoming the odds--doing better than predicted, given the factors working against them.
>
> The analysis shows that students in Bloomfield Hills and some other well-off, high scoring districts could be doing even better.
>
> The study also demonstrates how any straight-up comparison of MEAP scores is inevitably flawed. Consider: Only 49 percent of Detroit fourth-graders pass the MEAP math test, compared with 84 percent of Bloomfield Hills fourth-graders.
>
> Consider: Seventy-one percent of Detroit students are from families poor enough to qualify for a free or reduced-price school lunch; only 2 percent of Bloomfield Hills students qualify. (Van Moorlehem & Newman, 1998)

The *Free Press* staff conceded that educators and testing officials suggest that MEAP scores do mean something, particularly in terms of how well students have grasped Michigan's model curriculum and how well a school's subject area is aligned with that curriculum.

> But they don't tell you how effective the teachers are in a particular building, how challenging its classroom lessons are or how much progress its students have made. Comparing the scores "degrades our discourse about the nature of education, the nature of learning," said Hursh [University of Rochester]. "It enables us to ignore the real issues." (Van Moorlehem, 1998a)

Several articles in the series highlighted schools that were achieving scores at a higher level than predicted by the regression equations. *Free Press* staff identified special programs and school-centered efforts to improve achievement that they suggested

improved scores above what was expected or predicted given the school's demographics. It essentially addressed the question posed by the editors of *Quality Counts '98*, again: If one school can succeed under the worst conditions, with the neediest children, how can others be permitted to fail?

There were several problems with the analysis completed by the *Free Press* staff and resulting interpretations and discussions were misleading. The first stemmed from the combination of reading and math percent proficient. Reading and math scores were moderately correlated at best and a resulting combination of scores would certainly be more difficult to interpret. The use of percent proficient was also problematic and based on a dichotomous decision rather than the use of a central tendency indicator like mean scores. Some of the explanatory variables (predictors) were highly correlated; for example, percent eligible for free and reduced lunch was correlated at or above 0.80 with local unemployment, percent of single parent households, and percent of households where no one had a high school diploma. Out of the ten correlations among the five predictors, five were above 0.80. School funds per pupil was not highly correlated with any predictor nor with the outcome MEAP index (all correlations were less than 0.16). This suggested considerable multicollinearity.

More important, perhaps, was the resulting overall model fit for the three regressions. The R^2s for large, medium, and small districts were 0.62, 0.33, and 0.16. Certainly, for small districts where the regression equation explained only sixteen percent of the variance, resulting predicted scores were highly unrealistic. The use of such a regression equation to predict scores and rank schools accordingly was a serious abuse of statistical (un)certainty. The resulting regression equations are reported in Table 2. As

can be seen from Table 2, each of the three equations based on geographic location of the school district was estimated with different explanatory variables. Even for those variables used in the final models, not all were significant. In Rural districts, for example, only the percent of households with no one having a high school diploma was a significant explanatory variable. The variables that had an overall bivariate correlation with the MEAP index about 0.50 or greater included unemployment, median income, poverty indicator, free-lunch eligibility, and percent of households where no one had a high school diploma. Again, most of these indicators were intercorrelated at 0.80 or greater.

All of these indicators were essentially beyond the control of schools or school districts. It was unfortunate that the *Free Press* staff decided not to include any school district controlled indicators. It was also unfortunate that the *Free Press* staff did not independently identify schools with innovative instructional or other learning relevant programs and then check to see what their obtained and predicted MEAP scores were. Instead, they sought out schools that scored above what was predicted by poorly specified regression equations and identified programs within these schools that might be responsible for the achievement above what was expected (predicted).

Table 2
Regression Equations from Free Press Special Series on "Testing MEAP"

	Urban/Sub urban	Mid-size Towns	Rural Areas
Percent of households with children where income fall below federal poverty guidelines	.	-.115	-.163
Percent of students in district eligible for free or reduced-cost lunch	**-.310***	.	-.205
Percent of households in district with children where English is not spoken	.	.	-.068
Local unemployment rate	-.185	.	.
Median income of households with children		.	-.172
Percent of households with children headed by single parent	**-.231***	.	.
Percent of households with children where no one has a high school diploma	-.107	**-.335***	**-.196***
Per pupil revenue	**.135***	.	.
Foundation grant given to the district by state for basic expenses	.	.118	.095
Percentage of students new to the district that year	.	**-.249***	-.084

Note. The coefficients in the table are standardized beta-coefficients, reported to facilitate comparison of the relative impact of each predictor.
* $p < 0.05$.

Finally, the regression analysis based on MEAP scores at the school district level had lost a great deal of information by ignoring variation among schools within districts and variation among students within schools. A much stronger analysis could have been done using a hierarchical linear model with student level scores and school level indicators. In this way, the amount of variance due to schools could have been properly evaluated. Because of the interrelationships of the predictors and the use of multiple outcomes (four areas of achievement could have been used as distinct outcomes), the problem was actually a multivariate one, which might have been addressed better through

multivariate techniques or structural equation models. The above analysis essentially assumed that schools did not vary within districts, students did not vary within schools, that math and reading were measures of the same construct, and that the predictors were relatively unrelated. Of course, this was not true. More importantly, the researchers failed to include what the literature has suggested were achievement-relevant indicators or predictors that would support or inform the work of teachers, schools, and school districts. In fact, regression analyses such as these do little to provide useful information or guidance to the educators, policy makers, and the American public that the editors of *Education Week* called for earlier.

These two presentations of student achievement in Michigan were important to review because this was FRONT PAGE NEWS. Soon thereafter, on April 29, 1998, the headlines of the Detroit *Free Press* read: "Students, parents rebel against state test. In some districts, people don't even show up" (Van Moorlehem, 1998b, April 29). These were the analyses and investigations that received the greatest attention. These were the kinds of stories that made a difference to policy makers and legislators (Rodriguez, 1995). These were the kinds of stories that kept teachers up at night. What can a teacher do to improve achievement of his or her students with such information? Armed with the knowledge that poverty is the primary "predictor" of a student's achievement puts a teacher on the battle line with little more than a camera to take a picture of what has already been pre-determined. Although serious flaws can be identified in such analyses, serious attempts to find useful indicators of achievement have been few--at least serious attempts that have made it to the front page.

Specific Motivation

For the most part, statewide testing programs are tools for educational reform through accountability systems (much like the monitoring functions of NAEP or most national tests). Generally, the primary purposes for implementation of state-wide testing programs include achieving greater accountability for student achievement, motivation for schools to adopt state curriculum frameworks, incentive for schools to raise standards for achievement, and more generally to improve teaching practices and learning. Although some may argue that all of the reform purposes for employing statewide tests are ultimately to improve teaching and learning, little has been done to demonstrate such a result. The measurement of gains in learning has troubled the measurement community and educational researchers much longer than the existence of any state test.

More recently, in light of the myriad educational reform programs currently in place around the nation, educational researchers have made numerous attempts to understand statewide testing programs as reform initiatives. One only has to review the recent annual meeting program of the American Educational Research Association to estimate interest in statewide testing programs. Several states have undertaken serious attempts to understand the relationship between state-test performance and some classroom processes, most notably, instructional practices, but also assessment practices.

One glaring omission from this line of research is on the role of classroom-level assessment and the assessment practices and competencies of teachers. There is a growing literature on the topics of teacher classroom assessment practices and competencies. However, this literature has not attempted to link assessment practices to

student performance: Do the assessment practices of teachers make a difference? This literature and related literature was reviewed in Chapter II; however, the literature was reviewed and presented in the absence of a comprehensive theoretical framework for the assessment of school achievement. Cizek (1997) echoed the sentiments of Glaser and Silver (1994) who argued that "the theory underlying the assessment of school achievement is less explicit" (p. 400) than that regarding other purposes of testing. These issues were described as they arose in the literature review.

The closest thing to a comprehensive theory of assessment and achievement was a recent review of the literature by Brookhart (1997). She presented a theory about the role of classroom assessment in motivating student effort and achievement. The theory suggested that the classroom assessment environment "played out" in repeated assessment events through which a teacher communicated and students responded according to their perceptions. (More on this later.)

In the present study, athematics was the subject area chosen as the focus of the analyses for three primary reasons: (1) the national focus on science and mathematics within Goals 2000 and the Third International Math and Science Study (TIMSS); (2) the availability of data from the comprehensive mathematics assessment used by the Third International Math and Science Study, and (3) the comprehensive nature of the teacher and student background information available in the TIMSS database. Middle-school classrooms were chosen because of the importance of the transition period from grade school curriculum to high school curriculum (curricular differentiation is greatest at the high school level). It was assumed that middle-school students and their teachers have a

stronger sense of what constitutes mathematics and that during the years of early adolescence, subject-matter interests are solidified and differentiated.

Research Questions

The specific research questions for this project were derived from the above presentation of Michigan students' school achievement and the current interest in large-scale testing programs and classroom processes. They were an attempt to bridge a gap in the developing theoretical framework for the assessment of school achievement by providing absent links to classroom assessment practices.

1. What are the current assessment practices of mathematics middle school teachers?
   a. How frequently to teachers engage in assessment of their students?
   b. What types of assessment tools are used?
   c. How do teachers use assessment information for formative (instructional feedback) and summative (assigning grades) evaluation decisions?
2. How do students perceive the significance of feedback given by their teacher?
3. How do students perceive their self-efficacy regarding mathematics performance?
   a. Do students' attributions of control differ for mathematics?
   b. Do students' perceived potential for mastery differ for mathematics?
4. Are there differences in the above characteristics based on gender, language spoken at home, or other characteristics of classrooms (type of math class)?
5. What are the interrelationships of teacher assessment practices, feedback, student self-efficacy, student effort, and achievement performance?

a. Are teachers' assessment practices related to student achievement?

b. Are teachers' feedback and student self-efficacy related to student effort?

c. Given classroom assessment practices, and student self-efficacy, is student effort related to students' achievement performance?

6. Are there identifiable patterns among the assessment practices of teachers?

All of the relationships above were examined using the complete TIMSS assessment and classroom level database on teacher and student questionnaires.

CHAPTER II

Literature Review

"Classroom teachers are the ultimate purveyors of applied measurement, and they rely on measurement and assessment-based processes to help them make decision every hour of every school day" (Airasian & Jones, 1993, pp. 241-242). However, applied measurement specialists have repeatedly demonstrated the problems of teacher-made tests, item-writing errors, ill-defined rubrics for the scoring of alternative assessments, and other issues (for a review, see McMorris & Boothroyd, 1993). Few, however, are currently continuing the classroom-level assessment research agenda. And fewer make explicit connections between classroom assessment performance and student achievement or performance in large-scale assessment programs.

When Ebel was awarded the 1979 Award for Distinguished Service to Measurement, the citation commended him for his concern with developing the fundamentals of educational measurement, with disseminating these basic principles to teachers, and for his writing on practical test-construction and analysis, which upgraded the quality of achievement testing in the classroom. Ebel (1976) argued that "to measure achievement effectively the classroom teacher must be (a) a master of the knowledge or skill to be tested, and (b) a master of the practical arts of testing" (p. 76). Measurement specialists have suggested for years that the practical arts of testing should be covered in the teacher education curriculum. However, many teacher education programs seriously lack sustained training in classroom assessment (for a review, see Mehrens & Lehmann, 1991, pp. 50-53; Stiggins & Conklin, 1992, pp. 177-203).

Ebel (1982) also reiterated and addressed two long-standing problems of test construction identified by Lindquist (1936): (a) what to measure and (b) how to measure it. These two problems summarize many of the questions raised in the measurement literature and common concerns of teachers in terms of assessment design. "How is the construct of achievement defined in the classroom? If it is more than an objective, external measure (as might be obtained from a standardized achievement test), what else is involved and how is it measured?" (Brookhart, 1994, p. 298). For many teachers, effort, progress, and actual performance are important to different degrees. What to measure continues to be an important question. At the same time, how to measure those things a community decides are important is also a critical question.

Stiggins and Conklin (1992) reported that teachers spend at least one-third of their professional time on assessment activities which inform a wide variety of decisions made daily and directly influence students' learning experiences. However, there was little in the literature that provided evidence either theoretically or empirically for the connection between assessment activities and learning or achievement.

Learning

In order to place assessment in a context of importance, the primary assumption is that learning has occurred to some extent and assessment is a tool to measure the extent of learning. A teacher's assessment strategies may, in part, reflect their operating beliefs or theories about learning. The debate about whether or not teachers must understand learning theories continues (Phillips & Soltis, 1991). Whether their learning theories are implicit (from experience) or explicit (from knowledge and use of theories), learning

theories play a role in teachers' classroom practices. Assessment is, in part, an attempt to measure learning or achievement of objectives (whether stated or not). The evaluation of assessment practices is only meaningful in the context of learning.

Theories about learning have gone through several major challenges, each leading to more complex descriptions and processes. Learning theories have evolved from individual phenomena to social phenomena, from passive participation to active involvement of the learner.

Even the most modern theories of learning do not capture all that should or could be in the identification or definition of learning. Many have contributed a lifetime of research and reflection uncovering the dimensions of learning. Learning is mysterious and illusive. Any discussion of learning and the role of classroom practices, personal behavior, motivation, or interest, must be cautionary to the extent that a common ground is possible in terms of defining learning. Can it be simply what is measured? Is it particular to each teacher and his or her learning objectives, whether they are behavioral, cognitive, or affective? In practice, learning goals may or may not be specified. Assessments are directed at measuring these goals, to one extent or another. Inferences about learning and achievement of the goals are made based on assessment results. In practice, assessment and the resulting inferences are likely to occur without an explicit theory of learning.

Koellner, Bote, and Middleton (1998) argued that teachers hold conflicting views about the nature of learning and even about what good teaching looks like. They also suggested that it was the inconsistency in teachers' beliefs that motivated them to be innovative, to experiment in their classroom practices. Other measurement specialists, as

presented below, have suggested that classroom assessment practices should reflect the instructional practices in the classroom. If a teacher's orientation and view of learning is aligned with constructivist perspectives, then these should be followed through in the assessment tasks and activities student engage in.

Crowley (1997) summarized many of these arguments succinctly by describing the shift in mathematics education during the 1990s, during which time the mathematics community began redefining what and how they teach, as well as what and how they assess. "All too often, after creating an environment wherein students have, for example, used calculators and group work to investigate challenging and meaningful mathematical situations, we assess their learning through a standard in-class test" (p. 706). She argued that the classroom assessment strategies developed by teachers should reflect their instructional activities as well as instructional objectives and learning outcomes.

Classroom Assessment

Much of the literature regarding classroom assessment was in the form of professional development or inservice-related articles and books. Richard Stiggins at the Assessment Training Institute has been a leader in this literature (Stiggins, 1989, 1991a, 1991b, 1991c, 1993, 1994, 1995a, 1995b, 1997, 1998; Stiggins & Bridgeford, 1985; Stiggins & Conklin, 1992). Early on, the focus was to describe the ecology of the classroom assessment environment. Stiggins and Bridgeford (1985) surveyed 228 teachers from eight districts around the country and found that use of teacher-made objective tests increased across grades, from second to eleventh grade. Half of the teachers who used their own objective tests reported to be comfortable with that type of

assessment tool. Math and science teachers were more likely to use objective tests than were teachers of writing and speech. Use of published tests decreased across grades, but were most frequently used in math classrooms.

Teachers also rated their use of objective tests most highly for grading and reporting purposes. In fact, they rated teacher-made objective tests higher for all purposes (including diagnosis, grouping students, grading, evaluating, and reporting) than they rated published tests or performance assessments. The most common concern teachers reported about their objective tests focused on test improvement.

In one of the larger works, Stiggins and Conklin (1992) reviewed over a decade of research they conducted on classroom assessment. Studies of classroom assessment practices have focused on three types of decisions made by teachers, including (a) preinstructional decisions such as planning decisions, (b) interactive decisions made during instruction, and (c) postinstructional decisions.

Overall, Stiggins and Conklin (1992) argued that classroom assessments are not only "one of our indicators of educational *outcomes*, but these classroom assessments also are part of the very instructional *treatments* that produce the desired outcomes" (p. 2). After observing three sixth-grade classrooms for ten weeks, Stiggins and Conklin reported that "the reason prior assessment researchers had not delved into this arena must have been the fear of trying to come to terms with and make sense of this immense complexity" (p. 6).

Salmon-Cox (1980) reviewed the literature on classroom assessment practices and reported that teachers relied primarily on their own assessment activities for information on student achievement. Observations and classroom work were also important sources

of information. In a survey of high school teachers, 40 percent used their own tests, 30 percent used interactions with students, 21 percent relied on homework performance, six percent used observations of students, and one percent used standardized tests for information about the achievement of their students.

In their survey of 59 mathematics teachers, Stiggins and Conklin (1992) found that mathematics teachers relied more on teacher-made objective tests than on published tests or performance assessments, more so for grading purposes and less so for diagnostic purposes. There was an increasing concern about improving and managing teacher-made objective tests as grade increased, particularly for math and science teachers compared to writing teachers. Based on 290 journal entries from 32 of the teachers, secondary teachers used assessments to assign grades (36% of their assessments), evaluate student mastery of material (30% of their assessments), and diagnose individual and group needs (16% of their assessments). The most common assessment strategies included behavioral observations of students (29%), teacher-made tests (28%), and review of student work or products (27%).

Stiggins and Conklin also summarized the work of Shavelson and Stern (1981) who reviewed thirty studies of teacher decision making. Regarding planning decisions, teachers placed most emphasis on academic and ability variables; decisions made during instruction were based on social interaction with academic activities; and decisions made after instruction were clearly based on more than achievement results. Other characteristics of students noted by teachers included disruptiveness, work habits, consideration, group mood, and participation as well as motivation, attentiveness, and

attitudes. These student characteristics played an important role in teachers' planning of instruction as well as making evaluative judgments about students.

Stiggins and Conklin (1992) cited Natriello's (1987) review of classroom assessment research who concluded that

> studies of evaluation processes are found to be limited by the lack of descriptive information on actual evaluation practices in schools and classrooms, a concentration on one or two aspects of a multifaceted evaluation process, and the failure to consider the multiple purposes that evaluation systems must serve in schools and classrooms. (Stiggins & Conklin, p. 209)

A profile emerged regarding the assessment environment in most classrooms. Critical elements included the purposes of assessment, assessment methods employed by teachers, criteria used by teachers to select assessment methods, the quality of assessment tools, feedback, the characteristics of the teacher as the assessor, teachers' perceptions of students, and the assessment policy environment (Stiggins & Conklin, 1992).

McMorris and Boothroyd (1993) sampled 350 multiple-choice and constructed-response items from 41 mathematics and science teachers in grades seven and eight and found that among the mathematics teachers, computation items were most common.

Stiggins and Conklin (1992) ended with a set of unaddressed research questions based on over a decade of their own research and review of relevant literature. Among these questions: What was the assessment process like from the students' perspective? How did it affect learning and academic self-concept? Did it differ by sex, race, or social group? They argued that "we can only use assessments to help motivate, study and promote learning if we understand their effects from inside the learner" (p. 212).

Based on many of the suggestions made by measurement specialists who have conducted research and have written on measurement-related issues relevant for classroom assessment, assessment is a pervasive element in the classroom environment. Teachers review results of standardized tests, create tests of their own using various formats, evaluate completed student projects they developed or obtained from resource guides or textbooks, assign work to do be done outside of school. They ask questions, watch, listen, interview students, pose questions for solution by individuals or groups of students. Then, to one extent or another, they communicate their findings and evaluations to students, and in doing so, impact the learning process, including participation in instructional activities, self-selected learning activities, and assessment activities. Assessment has important functions, including communicating learning goals, particularly the subject-matter content and thinking processes valued by the teacher. Assessment impacts students by shaping study behaviors and general and academic self-concepts and self-efficacy; enabling self-adjustment; enhancing academic motivation and effort; and organizing and securing the storage of knowledge and skills. These suggestions come from both empirical and anecdotal evidence presented by dozens of authors of classroom assessment texts reviewed above and below.

Educational Measurement for Teachers

One of the earliest texts addressing educational measurement for teachers was Tiegs' 1931 text, *Tests and Measurements for Teachers*. "The principal function of measurement is to contribute directly or indirectly to the effectiveness of teaching and learning" (p. 3). He continued with a discussion of learning:

> Learning does not always parallel teaching; in fact, at many points and in many different ways, there are learning difficulties. Particular measurement devices which will reveal the exact location and the nature of these difficulties will aid the teacher in directing further learning. Test scores may be utilized to advantage in helping pupils visualize their objectives and goals in meaningful terms. (p. 11)

Tiegs best captured many concerns of today's measurement community when he suggested that "the major function of the informal objective test is the guidance of teaching. Testing is very definitely an element of the teaching cycle" (p. 254).

Since 1990, over two dozen texts have been published, all addressing educational measurement. Most of these texts covered the basics of educational measurement issues: a defense for the role of assessment; the role and specification of learning objectives; classroom test design issues; item writing; reliability and validity; test assembly, administration, scoring (item and test analysis) and reporting (including grading); review of various types of standardized tests; guides to selecting published test instruments; guides for assessing noncognitive domains; special issues in testing (special education, disability accommodations, legal issues); and others (Airasian, 1994; Carey, 1994; Chase, 1999; Cunningham, 1998; Ebel & Frisbie, 1986; Gallagher, 1998; Gredler, 1999; Hanna, 1993; Hopkins, 1998; Kubiszyn & Borich, 1996; Linn & Gronlund, 1995; McDaniel, 1994; McMillan, 1997; Mehrens & Lehmann, 1991; Nitko, 1996; Oosterhof, 1990; Oosterhof, 1999; Payne, 1997; Popham, 1990; Sax, 1997; Stiggins, 1997; Thorndike, 1997; Tindal & Marston, 1990; Ward & Murray-Ward, 1999; Weirsma & Jurs, 1990; Worthen, Borg, & White, 1993; Worthen, White, Fan, & Sudweeks, 1999). These texts varied greatly in terms of the use of research to support the materials within them. Some authors took more time to cover direct application for teachers by addressing them

directly in the text and supporting materials. Few explicitly described the links between assessment and learning, or student outcomes.

Nearly all of the authors discussed the uses of classroom assessment for (a) pretesting, (b) formative evaluation, and (c) summative evaluation. Pretesting is done to assess what students already know in order to plan instruction and is sometimes called readiness testing. Formative evaluation generally includes the informal assessment behaviors of the teacher, including questioning, observing, interviewing, and homework; in order to help teachers determine instructional effectiveness, group students, understand misconceptions or identify problems, and develop review materials and posttests. Formative evaluations can also help students understand important elements of the lesson as well as their own understanding and progress, although few authors discuss this. Finally, summative evaluations consists of posttests that are given after instruction is completed to help teachers determine their own effectiveness, evaluate student achievement or progress and assign grades.

In the cited texts, only a few of the most recent authors discussed the connections between assessment and learning, and even these discussions varied widely in their depth of presentation and implementation strategies for teachers. Most frequently, authors suggested that the assessment of students and their learning was a continuous process. However, these authors did not provide a theoretical or practical link for teachers to understand the integration of assessment, instruction, and learning. A commonly presented model was based on three elements: instructional outcomes or planning instruction, teaching strategies and activities or delivering instruction, and assessment strategies and activities (Airasian, 1994; Carey, 1994; Chase, 1999; Cunningham, 1998;

Ebel & Frisbie, 1986; Gallagher, 1998; Kubiszyn & Borich, 1996; Mehrens & Lehmann, 1991; Oosterhof, 1990; Sax, 1997). For most authors, this was the extent of the discussion on integrating assessment, instruction, and learning. "If testing and instruction are fully integrated, the content of each test is closely related to the instruction given students. And, subsequent instruction depends on how well students performed on prior tests" (Oosterhof, 1990, p. 217). This can be seen in Figure 1.

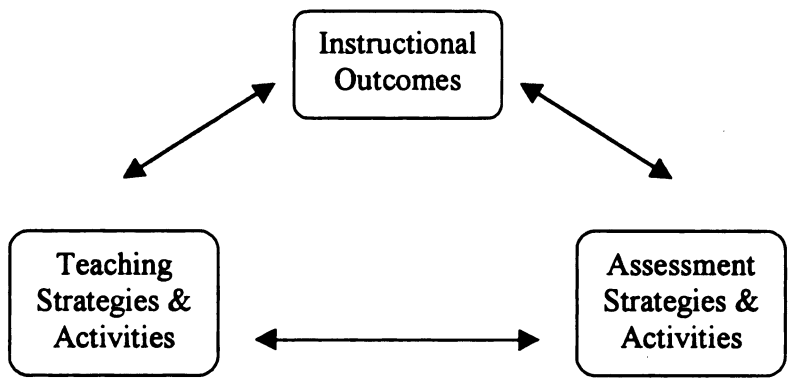

Figure 1. A model of instruction and assessment.

Some of the more practice-oriented authors still made references to the dichotomy between assessment and instruction. "Because assessment time 'costs' in instructional time, all of this (assessment and testing) must be done in an effective manner so that the investment in assessment yields maximally useful information" (Gallagher, 1998, p. 5). "It is important to distinguish between the instructional process itself and the pupil learning that results from that process. ...official assessment focuses attention primarily upon pupil achievement at the completion of that process" (Airasian, 1994). Generally, it

was a prevailing view that assessments that inform teachers about student achievement occur after instruction was completed, to determine instructional effectiveness, evaluate student progress, and assign grades. However, this prevailing view, that first we teach, then we test, is more myth than good practice (Stenmark, 1991). Instruction and assessment are actually more closely linked than most educators assume.

There were other authors who identified more important connections between instruction, assessment, and learning; that learning could occur during formative assessment and this process could be integrated with instruction.

> Going over tests constructed by the classroom teacher is an excellent technique for providing both feedback and a learning experience. Even the experience of taking the test itself facilitates learning. ... In summary, then, students learn while studying for the test, while taking the test, and while going over the tests after it is completed. (Mehrens & Lehmann, 1991, pp. 9-10)

Chase (1999), Ebel & Frisbie (1986), Popham (1990), Gredler (1999), and Wiggins (1998) supported these sentiments. Several authors also cited Stroud (1946), who suggested that "the contribution made to a student's store of knowledge by the taking of an examination is as great, minute for minute, as any other enterprise he engages in."

It is worth discussing Wiggins' (1998) ideas about educative assessment at this point. He, more than most authors, has taken the integration of instruction, assessment, and learning to a very practical level. In fact, he argued that "the aim of assessment is primarily to *educate and improve* student performance, not merely to *audit* it" (p. 7).

> First, assessment should be deliberately designed to teach (not just measure) by revealing to students what worthy adult work looks like (offering them authentic tasks). Second, assessment should provide rich useful feedback to all students and to their teachers, and it should indeed be designed to assess the use of feedback by both students and teachers. (Wiggins, 1998, p.12)

Some of the ideas presented were complex, but Wiggins also provided examples and worked through applications of his ideas. To describe some of these ideas more fully, he argued that "authenticity is essential, but authenticity alone is insufficient to create an effective assessment task" (p.30). Gredler (1999) also argued this point: the teacher's role is that of coach, guide, and facilitator; the student's role is that of learner and thinker in the subject area. So, according to Wiggins, the integration of assessment and instruction must involve the application of concepts and principles in the subject area to real-world tasks.

Wiggins also maintained that the assessment tasks assigned to students should not be considered instructional activities. The two can be integrated, but are not the same. The role of feedback was also critical. Not only should feedback after a student's performance be improved, but it should be provided during the assessment activity—concurrent with the assessment. "In other words, we must come to see deliberate and effective self-adjustment as a vital educational outcome, hence more central to how and what we test" (p. 43). Assessments that can educate and improve student performance must provide evidence of effective self-adjustment of the student. Assessments that are based on authentic real-world tasks provide the opportunity for students to receive feedback during the tasks, promote learning while engaged in the tasks, and encourage self-adjustment in responses or completion of the tasks.

The treatment of assessment in educational measurement texts varies a great deal. Although Wiggins' notion of educative assessments is compelling, it is far from current practice as presented by recent classroom assessment resources.

Teacher Competence in Classroom Assessment

One of the core propositions that first appeared in the policy statement of the National Board for Professional Teaching Standards, *What Teachers Should Know and Be Able to Do*, was that "teachers are responsible for managing and monitoring student learning," and that "teachers think systematically about their practice and learn from experience" (NBPTS, 1996, p. 16). Both of these spoke to competency in educational assessment of students. Several other authors have written extensively about teacher competency in educational measurement (Stiggins, 1991b; Plake, 1993; Plake, B. S., & Impara, J. C., 1997; Plake, Impara, & Fager, 1993), while others attempted to develop instruments to assess teacher competency (Stiggins, 1992, 1993; Zhang, 1996).

The American Federation of Teachers, National Council on Measurement in Education, and National Educational Association (1990) developed a set of standards for teacher competence in assessment. These organizations intended to provide a guide for teacher educators, a self-assessment guide for teachers, a guide for workshop instructors, and "an impetus for educational measurement specialists and teacher trainers to conceptualize student assessment and teacher training in student assessment more broadly than has been the case in the past" (p.1). These standards are outlined in Table 3.

At first glance, these standards appear overwhelming. With all of the tasks and responsibilities of the classroom teacher regarding content expertise and classroom management skills, how could any single teacher be competent in all of these additional tasks? The first three standards are the most frequently mentioned in classroom measurement textbooks and likely the critical elements, improving the effectiveness of

the following four standards. There have been several studies of teachers' competence in assessment development, some of which were mentioned earlier.

Table 3
Standards for Teacher Competence in Educational Assessment of Students

1. Teachers should be skilled in *choosing* assessment methods appropriate for instructional decisions.
2. Teachers should be skilled in *developing* assessment methods appropriate for instructional decisions.
3. Teachers should be skilled in administering, scoring and interpreting the results of both externally-produced and teacher-produced assessment methods.
4. Teachers should be skilled in using assessment results when making decisions about individual students, planning teaching, developing curriculum, and school improvement.
5. Teachers should be skilled in developing valid pupil grading procedures which use pupil assessments.
6. Teachers should be skilled in communicating assessment results to students, parents, other lay audiences, and other educators.
7. Teachers should be skilled in recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information.

McMorris and Boothroyd (1993) evaluated 350 multiple-choice and constructed response items from 41 mathematics and science teachers in grades seven and eight. Of those examined, 35 percent of the constructed-response and 20 percent of the multiple-choice items contained flaws. They also reported that among mathematics teachers, computation items were most frequently used. The overall quality of a teacher's test was also related to measures of the teacher's measurement competence. Finally, they argued that a test and its interpretation could affect students' attitudes about a class, the teacher and the subject matter. Potentially, assessment practices and skills have far reaching effects.

Learning, Achievement, and Assessment

Cizek (1997) provided a framework for understanding the uniqueness and interrelationships of learning, achievement, and assessment. Regarding learning, he suggested that definitions of learning have changed subtly to exclude any reference to student behavior and center on cognitive change exclusively. He referenced the works of Wittrock, T. L. Good, and Brophy. With respect to achievement, he suggested that since observable performance was excluded from the definition of learning, achievement must be defined without regard to learning. So, any notion of achievement must include some aspect of performance or behavior. Again citing Good and Brophy (1986), "the performance potential acquired through learning is not the same as its reproduction or application in any particular performance situation" (Cizek, p. 4). Noting that achievement is a "fallible representation or indicator of learning" (p. 4), Cizek argued that learning is not necessary for achievement. Indicators of performance generally can be ranked or certified whereas cognitive reorganization (learning) cannot. In fact, "because the relationship between learning and achievement is not direct, it serves to highlight the inferential nature of all assessment" (p. 4).

Cizek (1997) presented a set of conditions desirable for an appropriate definition of assessment. First, it should be applicable to current and future conditions, formats, and contexts; a generalizable definition is preferable. Second, a definition should enhance the role of assessment in instruction. Third, a definition should suggest that assessment serves rather than drives instruction. And fourth, a definition should include educational processes that promote the welfare of all students. Based on these considerations and the conceptual work of many other researchers, Cizek proposed the following definition:

> *assessment* \uh ses' ment\ (1) v.t.: the planned process of gathering and synthesizing information relevant to the purposes of (a) discovering and documenting students' strengths and weaknesses, (b) planning and enhancing instruction, or (c) evaluating progress and making decisions about students. (2) n.: the process, instrument, or method used to gather the information. (p. 10)

However, the point was well made that there exists a blurring between types of assessment, assessment for formative purposes versus assessment for summative purposes, as well as assessment that is distinct from instruction versus assessment that is integrated with instruction.

Assessment, Effort, and Achievement

A theoretical framework was recently offered that integrated two literatures: classroom assessment environments and social-cognitive theories of learning and motivation (Brookhart, 1997). The theory made explicit connections between the role of classroom assessment practices in motivating student effort and achievement. Brookhart defined a classroom assessment event in terms of the instruction given based on learning and assessment tasks and feedback provided to students, students' perceived task characteristics and their own perceived self-efficacy, students' effort, and their achievement. A version of this model adapted from Brookhart is illustrated in Figure 2.

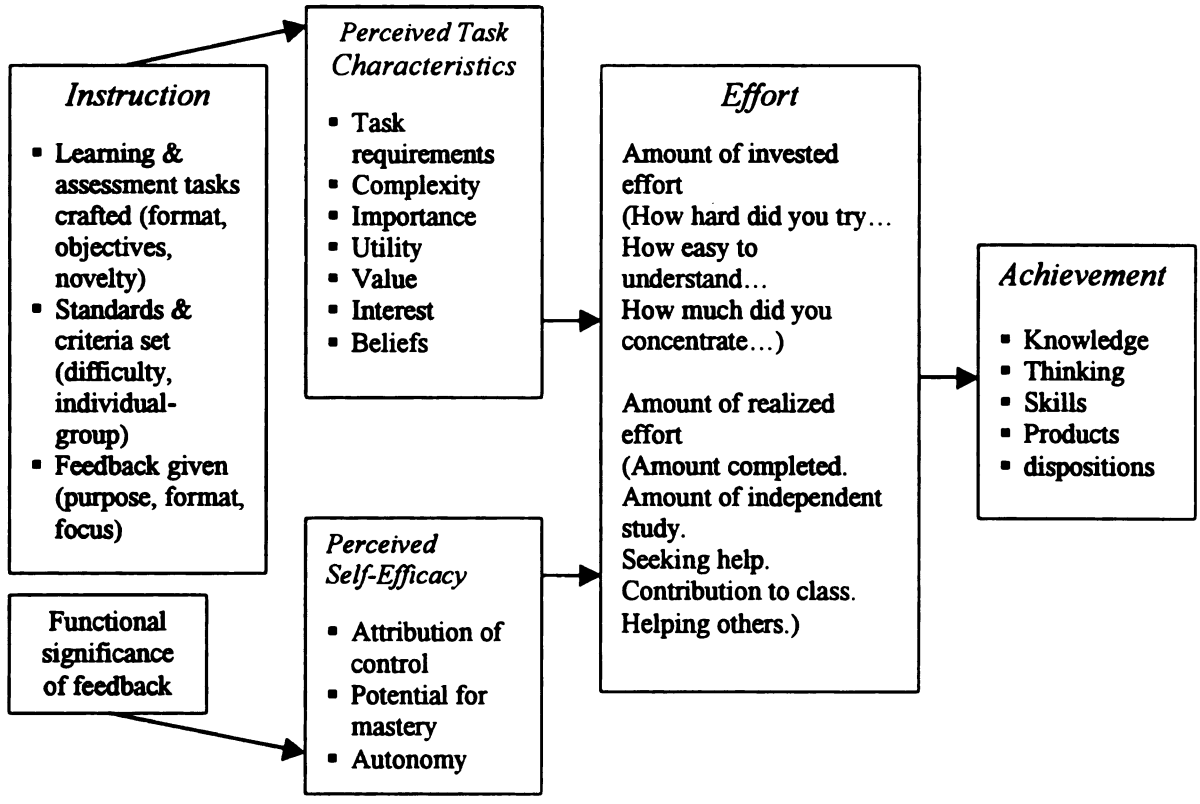

Figure 2. Model of a framework for investigating classroom assessment events.
Note. Adapted from Brookhart (1997).

A classroom assessment event could be described as a discrete set of objectives and assessments of whether the objectives were met. Brookhart (1997) argued that "the constitutive aspect of a classroom assessment event is its presentation of a task, activity, or set of tasks and activities where expectations are communicated and assessment is perceived" (p. 167). The perceived task characteristics differ for each student—different students perceive the same task differently. The functional significance of feedback can be perceived as informational or controlling and is determined by how the student experiences the event. Perceived self-efficacy includes "the student's belief or conviction that he or she can master the material, accomplish the task, or perform the skill that the assignment requires" (p. 173). The amount of invested mental effort includes the non-

automatic rehearsal of material, where realized student effort includes overt activity. "This theoretical framework should be able to predict the role of classroom practices in motivating student academic effort and achievement ... and is amenable to empirical testing" (Brookhart, pp. 161-162).

Brookhart argued that classroom assessment theory has several implications: (a) emphasis on raising classroom assessment quality; (b) use of a variety of student performances, particularly those meaningful to students; and (c) active involvement of students in the assessment process. According to Brookhart, teachers can use assessment tasks to communicate the classroom assessment environment to students and influence their effort and achievement.

Others have attempted to describe the relationships between assessment and effort explicitly. Camp (1992) suggested that assessment activities should exemplify worthwhile learning experiences; be based on meaningful tasks; integrate knowledge and skills; be flexible over extended periods; and occasionally provide opportunities for peer collaboration. Activities that encourage students to be responsible for their learning and understanding appear to improve motivation and effort.

In reviewing the work of Ames and Archer (1988) and Eccles and Midgely (1989), Blumenfeld, Puro, and Mergendoller (1992) argued that teachers' feedback, accountability, and evaluation practices affect students' motivational orientation, whether they are motivated to learn or simply perform.

> Students' expectancies for success are increased when teachers: (a) hold students accountable for learning and understanding—not just for getting right answers, (b) give students the freedom to take risks and be wrong, (c) stress improvement over time, (d) minimize comparison with others, (e) minimize competition, and (f) use private rather than public evaluation. (pp. 209-210)

Ward and Murray-Ward (1999) affirmed the role of student characteristics. "The motivational techniques, learning activities, content appropriateness, and management of consequences should match the person inputs (the components students bring to school which impact learning outcomes--cognitive and noncognitive)" (p. 323). Their view was illustrated in a flow-chart reproduced in Figure 3 below. Instructional factors and student inputs both affected student effort and performance, with an additional role for consequences not included explicitly in Brookhart's model.

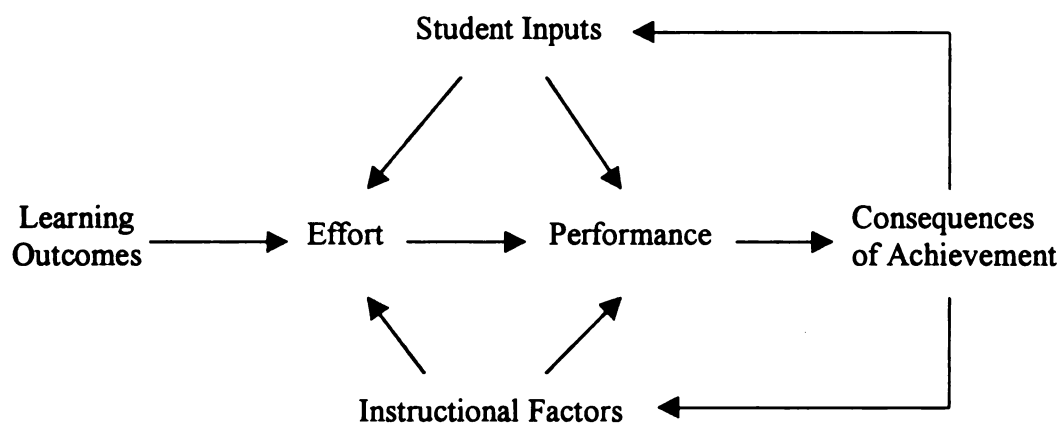

Figure 3. The relationship between student inputs, instructional factors, performance, and consequences.

Based on the considerations of the above literature, a modified model of classroom assessment practices and students' perceptions, effort, and achievement was proposed for this study. The Brookhart (1997) model was an "event" model, looking at the interaction of classroom practices, student perceptions and effort, and achievement within a given classroom assessment event. The model adopted for this project was

considered a generalization of the Brookhart Model based on a more general use of the literature. The generalized model is described in Chapter III (Methods). One additional consideration included the role of homework as a dimension of assessment practices.

Homework

Homework is an important yet controversial aspect of classroom assessment. It has often been one of the first areas targeted for improvement of student outcomes (Cooper, 1989). Cooper completed a meta-analysis on decades of experimental and quasi-experimental research on homework. Within that work, he synthesized 50 correlations between time spent on homework and achievement (33 correlations), grades (7 correlations), and attitudes (10 correlations) from 18 studies. The average correlation, weighted for sample size of the study, was $r = 0.186$, with a 95 percent confidence interval of $r = 0.180$ to $r = 0.192$.

Moderating variables were found that also influenced the size of the correlation reported in a study. The size of the correlation between time spent on homework and the outcomes was positively related to the year the study correlation was reported, suggesting that stronger correlations have been reported more recently than in the past. Studies conducted at the national level rather than state or local levels reported larger correlations. This may have resulted due to range restrictions either in homework time variability or achievement variability. Studies done in mathematics reported the strongest correlation ($r = 0.22$ on average). Cooper suggested that subjects involving long-term projects, integration of multiple skills, and creative use of outside resources (e.g., social studies) result in smaller relationships between homework and achievement than those

subjects involving rote learning, practice, or rehearsal (e.g., mathematics). Standardized tests ($r = 0.18$) and grades ($r = 0.19$) resulted in larger correlations with time spent on homework than outcomes related to attitudes ($r = 0.14$). Finally, the largest moderator effect was due to grade level. Studies involving students in high school grades 10 to 12 had a moderate correlation ($r = 0.25$), grades 6 to 9 had a small correlation ($r = 0.07$) and grades 3 to 5 had a nearly zero correlation ($r = 0.02$). An interaction between grade level and subject matter was also found significant. Specifically relevant for this project, the average correlation for high school students in mathematics classes was $r = 0.25$.

Cooper (1989) made important concessions in interpreting these final correlations. The correlations

> cannot be interpreted as demonstrating a causal effect of homework on academic achievement or attitudes. It is equally plausible, based on these data alone, that teachers assign more homework to students who are achievement better or who have better attitudes, or that better students simply spend more time on home study. (p. 100)

In addition, Cooper found evidence suggesting that the relationship between time spent on homework and achievement may be curvilinear for middle-grade students, where increases in the amount of time spent past 10 hours a week had no relationship to achievement.

Several studies have been conducted since Cooper's synthesis. Walberg (1991) reported that on average, 8th grade students spend about 1 hour each day on homework. Reports from the National Assessment of Educational Progress have demonstrated an increase in the proportion of 8th grade students who reported doing homework (Anderson, Mead, & Sulllivan, 1986). Eighth grade students reported time spent on homework was significantly related to a composite achievement measure based on data from the National

Educational Longitudinal Study (NELS; Keith, Keith, Troutman, Bickley, Trivette, and Singh, 1993). Using High School and Beyond (HSB) data, Keith and Cool (1992) reported that the average amount of time spent on homework per week had a small but significant effect on high school seniors' level of achievement, and motivation had an indirect effect on achievement through its relationship with quality of instruction and the amount of academic coursework taken (all from student reports). They suggested that "students enrolled in a high-quality school and curriculum are more highly motivated by that curriculum. Students with high academic motivation take more academic coursework ... and do more homework ... and as a result, achieve at a higher level" (p. 215).

Most recently, Cooper, Lindsay, Nye, and Greathouse (1998) surveyed 82 teachers and 709 students and their parents from three school districts in grades two through twelve. They reported nonsignificant correlations between teacher reports of amount of homework assigned and classroom achievement. In addition, students' reports of time spent on homework were not correlated with achievement, but were correlated with teacher-assigned grades ($r = 0.17$).

Many of these correlational studies also had methodological problems. In most cases, when nested data were used, particularly in the large national databases (i.e., NELS and HSB), dependencies within classroom or school were ignored. It is likely that relationships between homework, as a classroom practice, and achievement, are likely dependent on classroom level or school level characteristics. This is a prevalent error in most of the research investigating classroom level characteristics and student level outcomes.

TIMSS Conceptual Model

The above literature review was completed independent of the review of the conceptual model for the Third International Mathematics and Science Study (TIMSS). The data used in this dissertation came from the TIMSS USA Database. Early on, the TIMSS project was conceived of as a study of educational opportunity. They conceptualized student learning as being influenced by psychological theories of individual differences and motivation, as well as sociological concepts including family background. Essentially, "this view recognizes that educational systems, schools, teachers and the students themselves all influence the learning opportunities provided and in fact all are part of the definition and parameters that frame the opportunities of individual students" (Schmidt, 1993, p. 29).

Through a revision of earlier models used in previous IEA studies and existing research literature, a conceptual model related to student factors was designed to guide the formulation of instruments for the TIMSS.

> The model ... suggests that student background, the student's own academic history, the economic and cultural capital of the family, the belief students have about how to succeed in science including their self concept, the social press created by peers and teachers which exists in the classroom for encouraging involvement in science and how students spend their time outside of school together influence the motivation and interest a student has to study science and mathematics coupled with the effort they expend. (Schmidt, 1993, p. 28)

This conceptual model was very similar to the model presented by Brookhart (1997). It can be seen in Figure 4. The role of effort and motivation as a moderator of achievement and performance was key to both the Brookhart model and Ward and Murray-Ward's model.

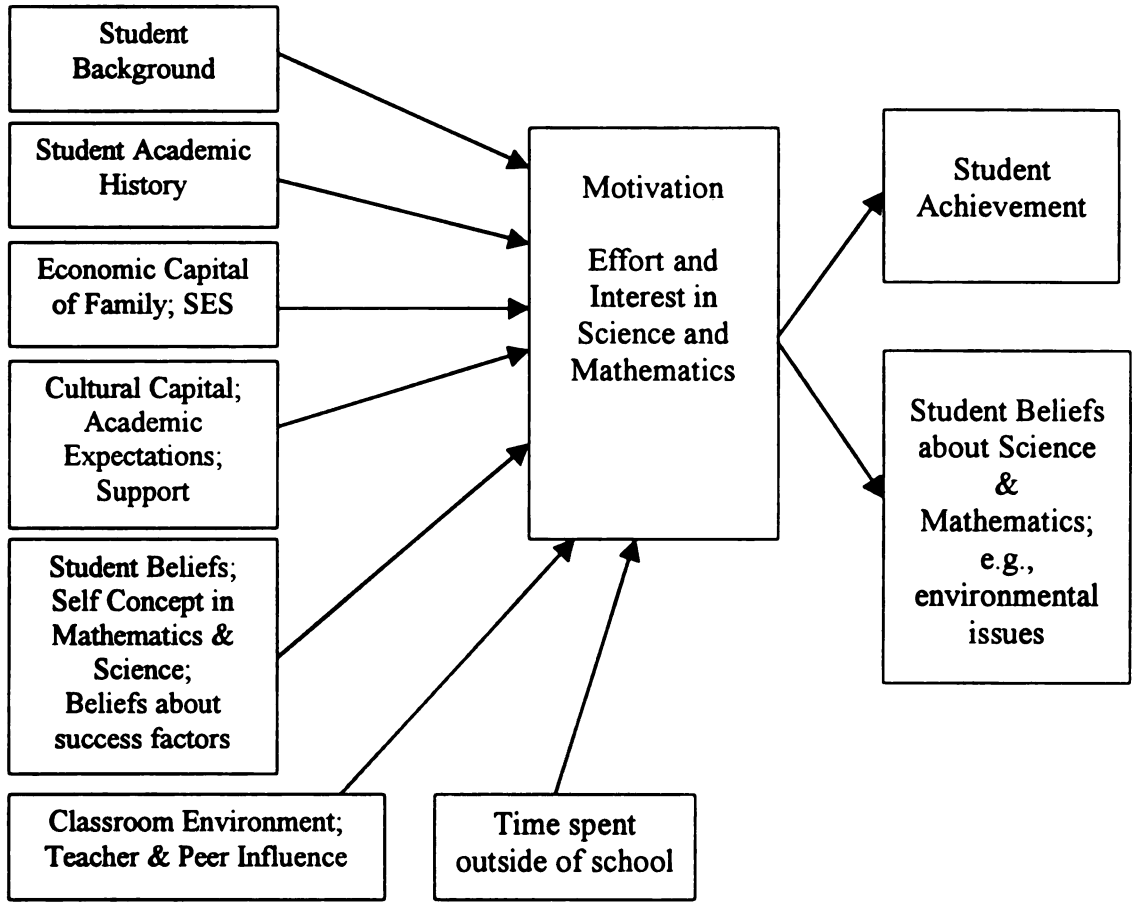

Figure 4. TIMSS Student Factor Conceptual Model.

Grades

Finally, although grades are not included directly in the analyses for this dissertation, they are a reality in classroom assessments and commonly the end-result or goal of summative evaluation practices of teachers. Estimates of how much tests and assessments comprized students' mathematics grades included anywhere from 25 percent to 80 percent (Thompson, Beckman, & Senk, 1997). However, grades are rarely unidimensional measures of performance. In fact, most teachers will admit that they

consider effort, improvement, and actual performance in the assignment of final grades. It could be argued that because of this, grades do not mean the same thing to all teachers, all students, and all parents.

Cizek (1997) stressed the distinction between assessment and evaluation and that "teachers' grading practices have been shown to be highly variable, and grades to be somewhat unreliable indicators of student achievement" (p. 29). Stiggins and Conklin (1992) found that "grading is the single most regular and influential feedback activity conducted by classroom teachers" (p. 175), yet grading practices have frequently violated guidelines promoted by the measurement community (Ebel & Frisbie, 1986; Gronlund, 1985; Mehrens & Lehmann, 1991). Most commonly, teachers have included other student characteristics in addition to achievement, including how much students were capable of learning or their level of motivation and effort; and teachers have treated daily assignments intended for practice and formative purposes as summative results to be combined with summative assessments.

> It is possible to use grades and other assessment feedback developmentally, not only judging the work (e.g., "poor") but also explaining what the student needs to do better. Cognitive evaluation theory suggests that if students get feedback that helps them make progress, then motivation and control should increase. (Brookhart, 1994, p. 296)

This suggested that although grades should be based on classroom assessments that are summative in nature, their use should not necessarily be confined to a summative report of achievement. It appeared undeniable, as summarized by Cizek (1996), that grades provide the primary mode of communicating to students, parents, teachers, and others, important information about student progress and achievement but that in practice, grades fall short of these expectations.

Grades present an incredibly difficult challenge as measures of performance. Validity of classroom assessment is blurred because of the "on-demand" nature of grades. Grades are due at the end of a grading period and this blurs interpretation and use (Messick, 1989). The reality that most teachers confound performance and effort in the assignment of grades has also been reported by many researchers (for a review, see Brookhart, 1994). To what extent can grades be reliably used in quantitative investigations of classroom performance?

Brookhart (1994) argued that the advice of the measurement community for grading has not taken "into account the teacher's need to manage the classroom and motivate students" (p. 299). Cross and Frary (1999) also cautioned against "abandoning or adopting recommended practices selectively depending on the values of each teacher and the culture of each classroom" (p. 69). They agreed with the suggestions of Troug and Friedman (1996), who suggested that the measurement community demonstrate the benefits of sound measurement practices by working with teachers in their classrooms.

Cross and Frary (1999), in their comprehensive study of teachers' attitudes and practices regarding grading and their students' attitudes as well, confirmed other reports of "hodgepodge" grading practices. They also reported that students not only confirmed such practices but supported them as well. Given such conditions, they finally suggested that measurement specialists should work harder to communicate with a broader audience regarding best practice.

These ideas about grading have parallels in terms of other classroom practices, including assessment. Communication within the measurement community is important as well, to clarify and uncover the nature of best practices. However, if measurement

specialists are unable to effectively communicate these findings to measurement practitioners (e.g., teachers), their work is futile. Methods to effectively balance costs and benefits and to demonstrate the utility of best practices must be communicated in a way that acknowledges the conditions of the classroom. These ideas are amplified in the next section.

Assessment in the Context of Educational Reform Efforts

As suggested in the introduction, accountability systems in place at every level of the educational system have as one general goal the improvement of educational achievement. How these improvements are realized is a critical question. However, an equally important question is where these improvements are realized. Many accountability systems include curriculum frameworks and large-scale tests at the state level (Sheilds, Corcoran, & Zucker, 1994). Such tests have several purposes, as discussed earlier, including the reform of classroom instructional and assessment practices. Professional educational organizations have also promoted classroom assessment standards to reform practice. Others have promoted special approaches or specific practices to ultimately improve achievement. Some have even argued that in order "to change practice, it is necessary to change practitioners, the classroom teachers" (Zucker, Shields, Adelman, & Powell, 1995, p. 21).

Many of the promoted reforms have included authentic assessment approaches (Newmann, 1997), educative assessment (Wiggins, 1999), the use of portfolios (Gitomer & Duschl, 1994), and others. Thompson, Beckman, and Senk (1997), in writing to mathematics teachers, suggested that assessment standards have been developed to

improve the mathematical abilities of students--to provide more information to the teacher and students. They argued that the National Council of Teachers of Mathematics assessment standards (described below) were founded in a shift toward assessing growth in mathematical power and away from evaluation of specific knowledge and isolated skills. The standards promote a shift toward more complex assessment tools and use of multiple sources of information and away from reliance on brief quizzes and chapter tests.

The National Council of Teachers of Mathematics (NCTM) has adopted a set of standards for assessment in mathematics classrooms in their Principles and Standards for School Mathematics (NCTM, 1998). NCTM has adopted a set of guiding principles to direct their work, of which one focused on assessment: "Mathematics instructional programs should include assessment to monitor, enhance, and evaluate the mathematics learning of all students and to inform teaching" (NCTM). The National Science Foundation attributes important similarities in how states view good mathematics education to documents such as the standards developed by NCTM (Zucker, Shields, Adelman, & Powell, 1995).

The evaluation standards developed by NCTM were presented separately from the curriculum standards, not because they believed that evaluation should be separated from the curriculum, "but because planning for the gathering of evidence about student and program outcomes is different."

> A common response to the challenge of the Standards is, "Yes, but who will change the tests?" Although pragmatic, this question shifts responsibility for change away from the individual to some unnamed higher authority. More productive--and more likely to make the vision embodied in the Standards a reality--are such responses as, "In what ways does the curriculum need to be changed?" "How best can these changes be

> made?" "How will we know when we have reached the Standards?" It is in the answers to these questions that the role of evaluation emerges as a critical component of reform. Evaluation is a tool for implementing the Standards and effecting change systematically. The main purpose of evaluation, as described in these standards, is to help teachers better understand what students know and make meaningful instructional decisions. The focus is on what happens in the classroom as students and teachers interact. Therefore, these evaluation standards call for changes beyond the mere modification of tests.
> (NCTM, http://www.enc.org/reform/index.htm)

These evaluation standards proposed that student assessment should be integral to instruction, that teachers should use multiple methods of assessment, and that teachers should assesses all aspects of mathematical knowledge and its connections. The curriculum and evaluation standards (NCTM, 1989) are currently under revision. The version currently in use contains ten standards for evaluation, which include (1) alignment with the curriculum, (2) the use of multiple sources of information, (3) the use of appropriate assessment methods and instruments; and several aspects of mathematical knowledge that should be assessed, including (4) mathematical power, (5) problem solving, (6) communication, (7) reasoning, (8) mathematical concepts, and (10) mathematical procedures.

The integration of assessment and instruction, the use of multiple methods of assessment, and the broad coverage of skill and knowledge in assessments that are driven by objectives identified by the curriculum are all messages supported by the measurement community. All of these ideas have been recommended in the majority of the textbooks on classroom assessment reviewed above. Whether or not mathematics teachers currently assess students in these ways, whether or not explicitly adopting the NCTM standards, can be evaluated in part from the TIMSS. Primarily, TIMSS provided

information on a variety of assessment tools and activities teachers have engaged in, but no information was available regarding their integration of instruction and assessment (except for the role of assessment information in planning future lessons) and no information was available on the extent to which teachers' assessments covered the full range of skills addressed in their curriculum or any specific learning objective.

Some recommendations have come from the TIMSS study to date, primarily concerning the relationship between curriculum and achievement (SciMathMN, 1996). Based on analyses of curriculum and achievement, the Minnesota TIMSS project had three recommendations regarding the improvement of "how we measure." Minnesota was sampled as a mini-nation in addition to their inclusion in the national sample. Their recommendations focused on the relationship between the Minnesota graduation standards and the statewide tests, suggesting that the statewide test include more demanding items (e.g., open-ended questions or student-constructed response problems). They also recommended that the curriculum, instruction, and assessment practices be analyzed to insure all students, particularly those traditionally under-served in mathematics and science education, receive the opportunity to learn. Finally, they recommended that funding and incentives be implemented for local alignment with statewide standards and assessment.

The recommendations from the Minnesota TIMSS study appear laudable. However, the lack of such specific findings and subsequent generalized conclusions seem too weak to ensure strong policy development in the area of mathematics and science assessment reform. Their statement regarding the inclusion of more demanding items on the state test is too simplistic, in effect assuming that constructed-response items are more

demanding with no regard for cognitive objectives. The idea that assessment practices should be evaluated to insure that all students received adequate opportunity to learn state objectives was also limited in that no evidence of tracking or differential assessment practices based on tracking was available in the TIMSS results.

Unfortunately, much of the analyses completed for the Minnesota TIMSS project has been exploratory, unguided by a theoretical framework. Similarly, many of the earlier evaluations of assessment practices of teachers have also been exploratory. As described by others (e.g., Cizek, 1997; Glaser & Silver, 1994), the theory underlying assessment of school achievement is less explicit than theories regarding other purposes of testing. The lack of explicitness inherent in investigations regarding classroom assessment has lead to generalizations that are often evaluated outside of meaningful contexts. Without grounding evaluations of classroom assessment in appropriate contexts, their impacts on achievement will be difficult to infer.

Toward a Theory of Classroom Assessment

As suggested above, much of the work done in classroom assessment research has been exploratory, without a strong theoretical framework on which to base hypotheses. However, several measurement specialists have considered what such a theory may look like or consist of, and for what purposes a theory of classroom assessment may be put to use.

Brookhart (1997), in her theoretical framework for the role of classroom assessment, has presented the closest model to a theory of classroom assessment. "Classroom assessment theory has implications for how teachers design and use

classroom assessment and for what teacher educators must prepare teacher to do" (p. 178). Such a theory could potentially support current efforts to raise the quality of classroom assessment practices because it is likely to impact more than the validity of the resulting information, an important but not exclusive aspect of quality assessments.

The need for a prescriptive theory of instructional test design has been argued by Nitko (1989), who suggested that such a "theory would predict which test design would be most appropriate in a particular instructional procedure under given instructional conditions and for specified instructional outcomes" (p. 417). The elements classroom assessment must address can be obtained from knowing the instructional decisions that are to be based on the resulting information.

> When a teacher (or other instructional developer) is in the process of deciding which instructional method is best for bringing about the desired changes in specific types of students and for a specific course's content, the teacher or developer should also be deciding on the best testing procedures for bringing about these changes. (Nitko, 1989, p. 448)

Because of the complex nature of the demands faced by teachers in diverse classrooms, prescription may never result from any comprehensive theory of classroom assessment. However, to the extent that teachers understand the contingencies inherent in the connections between content, student characteristics, and instructional decisions, a teacher should have available a repertoire of assessment practices to meet those contingencies. First, however, it is important to uncover the nature of these contingencies and the complex nature of interactions between teacher decision requirements, instructional activities, course content, student characteristics, and elements of classroom assessment practices.

A theory of classroom assessment should be able to prescribe elements of instructional test design to meet the challenges faced by teachers and should also inform teacher education programs and professional development activities.

CHAPTER III

Methods & Procedures

Research Design

The primary research orientation was quantitative. The study was primarily a correlational study, examining existing conditions of several classroom characteristics and their impacts on student achievement. The unit of observation (or the subject of the study) was the student. However, teacher and classroom level data were available, making the resulting data set hierarchical where students were nested within classrooms.

In the language of quasi-experimental designs, the classroom assessment practices of the teachers were the treatment conditions. Classrooms were non-equivalent groups of students that differed in many ways (some more than others), in addition to the characteristics of their teacher's assessment practices. The outcome was student performance on the TIMSS mathematics assessment.

This study employed the position of Cook and Campbell (1979) regarding any future argumentation of causality. Particularly relevant were the following arguments: (a) complex systems of causality are contingent on many conditions and causal laws operating in such systems are fallible and probabilistic; (b) effects follow causes in time and may be instantaneous; (c) effects in such complex systems of causality can be the result of multiple causes; (d) causality is difficult to identify in most field research involving open systems where there likely exists other mediating causes involved in the absence of the cause of interest; (e) some causal laws work in reverse where cause and effect are interchangeable; and (f) the manipulation of a cause will result in the

manipulation of an effect. It is this last point that drives much of the applied research in education; however, the actual manipulation of a cause is rarely achieved in field research, particularly correlational research such as this. The researcher designs a study (in whatever form) to gain an understanding of how the complex organization and system of education and its integral processes work with the expectation that specified educational outcomes are potentially controllable to some degree. More generally, the researcher hopes to at least learn more about which processes (more or less in the control of educators) hold promise for influencing educational outcomes.

It was reasonable to expect causality to flow in two directions in the consideration of classroom practices and their effects on achievement. For instance, initial information teachers obtain on students' prior experiences (perhaps from pre-test scores) inform their instructional strategies. A teacher starts a lesson based on what the students know coming into the lesson. Simultaneously, students perceive what are the important skills and information to learn based on the teachers' practices (instructional and assessment tasks). Student achievement is affected by student perceptions and classroom practices, which were initially informed by prior student achievement. This cycle could be replayed throughout a school year, but limited to the degree that teachers actually make use of achievement information in their instructional and assessment practices.

Based on an extensive review of the literature and the frameworks provided by Brookhart (1997) and the TIMSS conceptual model (Schmidt, 1993), a hybrid model was assessed. The model is presented as a graphic representation of a path model. The components of the measurement model are illustrated in Figures 5 to 8 while the structural model is illustrated in Figure 9. In addition, the model was evaluated as a

hierarchical linear model as described below. The model hypothesized the primary relationships and guided the assessment of the measurement properties of the constructs under investigation. The hierarchical linear model allowed for appropriate accounting of the nested nature of the data and included additional features such as demographic information of the students; however, it ignored the measurement error in the constructs as defined below. These methodological issues will be described more fully in later sections.

Finally, learning was not explicitly represented in the models used in this study. In fact, learning was also absent in most of the models presented earlier. A broad perspective on learning has been adopted in this project. It was not assumed that there was a particular kind of learning taking place in mathematics classrooms. For some children, in some instances, and for certain topics and activities, learning likely encompassed behavioral, affective, cognitive, and-or social elements that may have been achieved through didactic interactions or through individual or social construction. Most importantly, learning was occurring through the exposure to lessons and evaluative tasks, to the teacher and other students. These elements, in combination, provided students with learning opportunities as defined earlier by Schmidt (1993). These opportunities manifested themselves in terms of student's mathematics self-efficacy and effort, and in turn, their achievement.

Modeling

The proposed models were framed in terms of structural equation modeling. Structural equation modeling provided a way to estimate several equations

simultaneously, accounting for measurement error in estimating latent traits and structural relationships among those latent traits. In the models as presented, the variables in boxes represent indicator variables for which observed responses were available. The model assumed that one or more latent traits, represented by circles in the model, were measured by the observed indicators; the latent traits exhibited themselves in the pattern of responses observed in the indicator variables. Each indicator or observed score was thus caused, in part, by the latent trait. In addition, each indicator was not entirely caused by the latent trait (common factor); the variance remaining, given the latent trait, was error variance (unique factor). In this sense, structural equation modeling (SEM) was a confirmatory technique. It required the specification of a model to be tested, a model that was optimally based on theory. Also, all of the paths in the model had significant directional influence, considering the influence of any preceding variables along a given path.

As an alternative, hierarchical linear model (HLM) approaches have been developed with accepted estimation procedures where effects are appropriately defined at various levels; HLM use has become common in educational research (Frank, 1999). Hierarchical Linear Models 4.03 (Bryk, Raudenbush, & Congdon, 1996a, 1996b) is recognized as a standard program for estimating hierarchical models (Bryk & Raudenbush, 1992; Kreft & DeLeeuw, 1998). Estimation using HLM relies on assumptions similar to multivariate multiple regression. Unfortunately, the measurement error inherent in constructs as measured by questionnaires is basically ignored in these linear models. However there are several unique strengths the are provided by the use of HLM, as described in a later section.

Identification and Measurement of Teacher Classroom Assessment Practices

Teacher classroom assessment practices were multifaceted and multidimensional. No measurement specialist has suggested that a single scale could be constructed to capture the essential aspects of a teacher's assessment practice. In fact, no scales that attempt to describe even pieces of such practices have been identified. There were instruments in use that help identify elements of assessment practices, however. For example, Stiggins (1998) developed a classroom assessment practices questionnaire (self-evaluation) in his work providing professional development activities to teachers through Assessment Training Institute. However, no attempt has been made to scale the results as a way of describing assessment practices on a continuum or categorically.

Teacher assessment practices were investigated for this project in two facets. Because of the prevalence of homework in secondary mathematics programs and because homework is often the first line of reform efforts for classroom practice, homework was examined as a unique and important facet of classroom assessment. Within the homework facet, there were two dimensions. Given that homework was assigned by teachers, the first dimension included the kinds of homework tasks that were assigned. The second dimension included the uses of the assigned homework as a facet of classroom assessment.

For the second facet of classroom assessment, all other assessment practices were included. However, these were multidimensional as well. For the purposes of this investigation, two dimensions were employed. The first described the types of assessments -- the tools used by teachers in their classroom assessment routines. The second included the uses of the assessment information -- what teachers did with the

information obtained through their classroom assessment routines. These were complex dimensions, particularly with regard to assessment tools.

So as can be seen in Figures 5-9, each element of the model was measured with multiple indicators. Classroom assessment practices (Figure 5) included both homework related practices and other assessment practices, where each was described by the tools employed and the uses of the resulting information as described earlier. Figures 6-8 include the items used to measure student effort, student self-efficacy, and the significance of feedback. Finally, Figure 9 illustrates the larger model, combining each of the teacher and student characteristics as measured.

The types of assessment practices at a teacher's disposal could be broken down further in many ways (e.g., traditional/nontraditional, objective/subjective, norm-referenced/criterion-referenced, etc.). Many of these categories are based on philosophical or pedagogical orientations, but selection of an assessment tool should ultimately be driven by the goals of the teacher in their cycle of setting learning objectives, designing instructional activities, and assessing student achievement, and to whatever degree these activities are integrated.

The TIMSS database (from the teacher background questionnaire) contained elements of these facets and dimensions. Each is described in turn. Appendix A contains frequency tables for each question and each possible response.

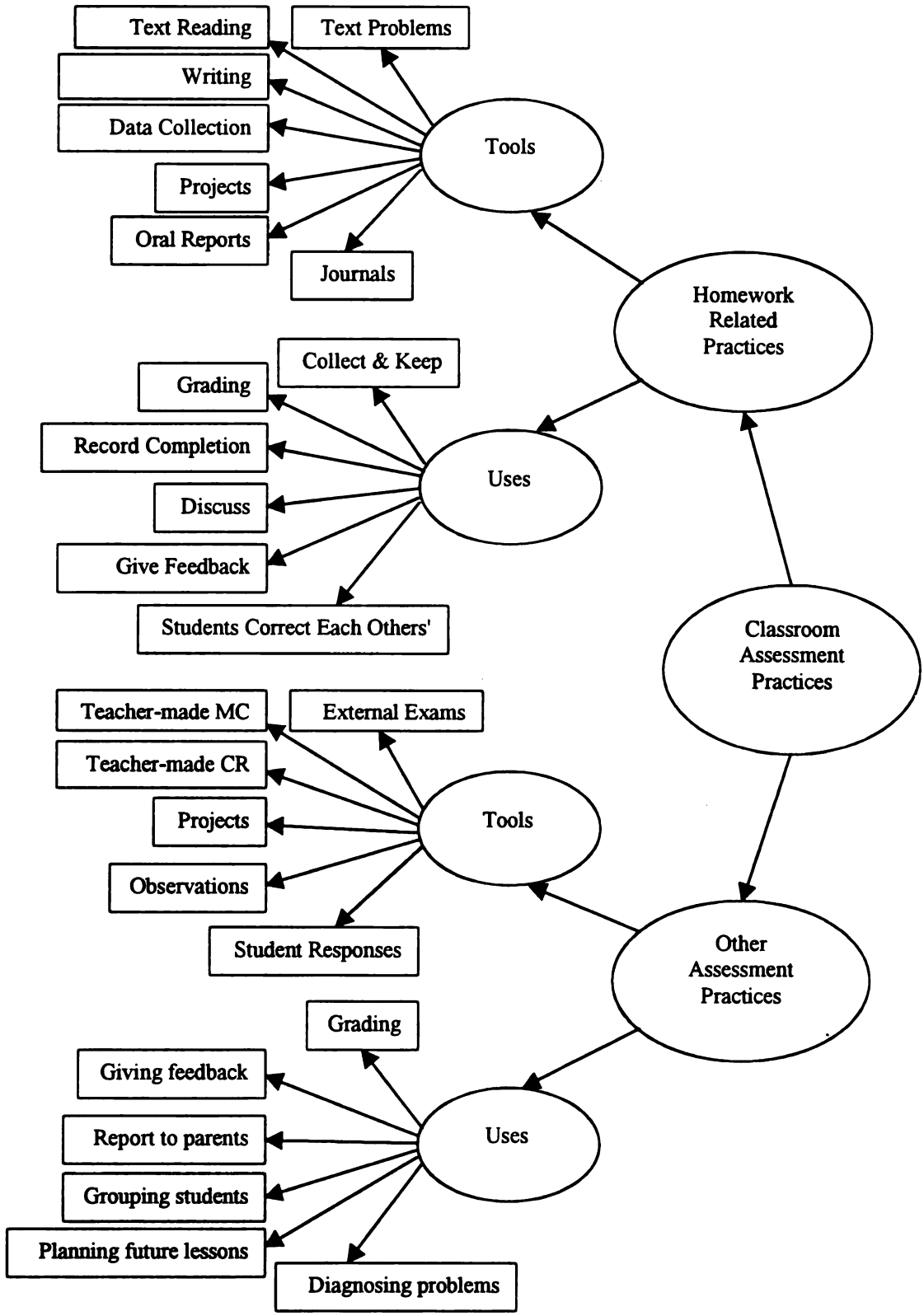

Figure 5. Measurement model for classroom assessment practices.

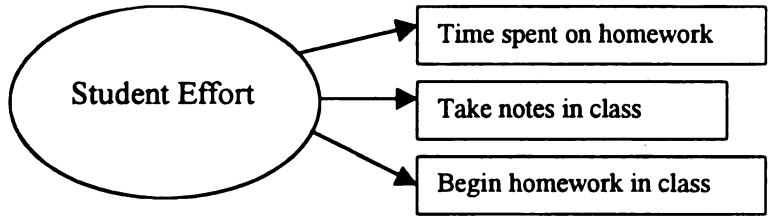

Figure 6. Measurement model for student effort.

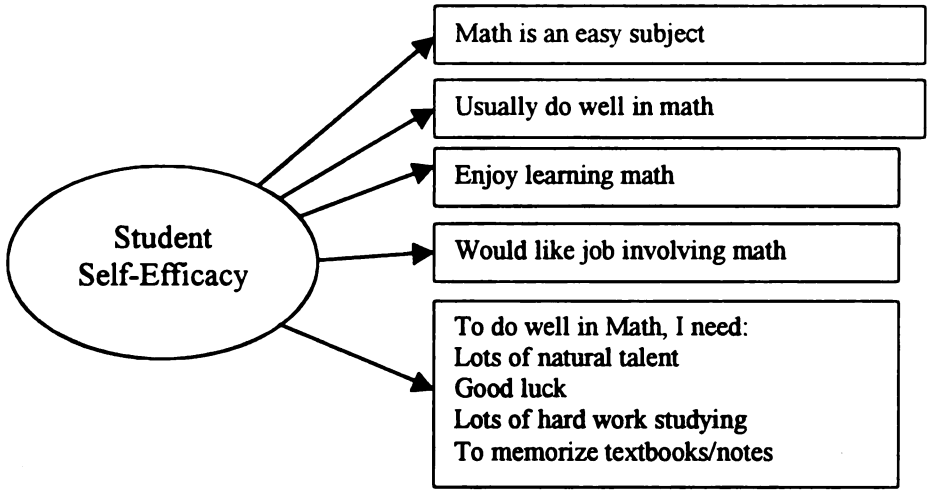

Figure 7. Measurement model for student self-efficacy.

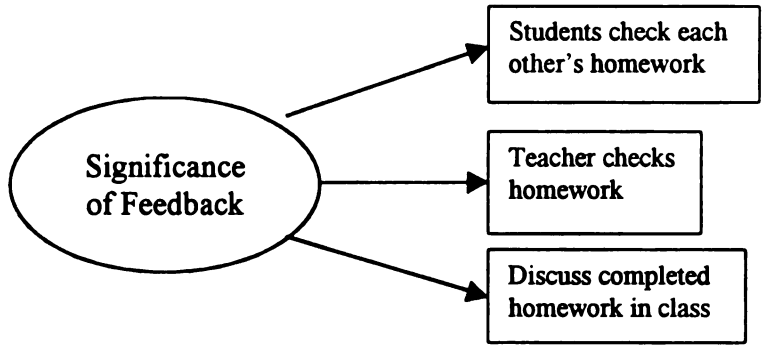

Figure 8. Measurement model for significance of feedback from students' perspective.

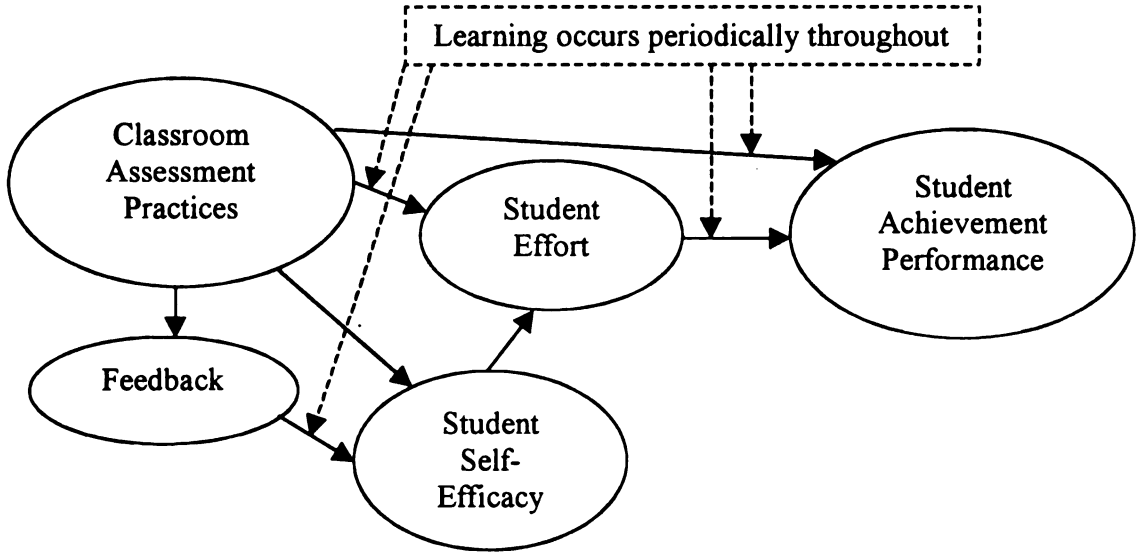

Figure 9. Structural model, illustrating the relationships among the latent traits.

The structural model shown in Figure 9 could also be configured as a hierarchical linear model, since it is, in one sense, a system of linear regression equations. Additional demographic variables pertaining to students (e.g., gender, use of English at home, mother's education level, and mother's expectations for mathematics achievement) could easily be accommodated in this configuration.

Student level effects:

$$\text{Achievement}_{ij} = \beta_{0j} + \beta_{1j}\,(\textit{Gender})_{ij} + \beta_{2j}\,(\textit{English})_{ij} + \beta_{3j}\,(\textit{Effort})_{ij} + \beta_{4j}\,(\textit{Self-Efficacy})_{ij} + \beta_{5j}\,(\textit{Mothers' Education})_{ij} + \beta_{6j}\,(\textit{Mothers' Expectations})_{ij} + r_{ij}$$

Classroom level effects:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\,(\textit{Class-Type})_j + \gamma_{02}\,(\textit{Feedback})_j + \gamma_{03}\,(\textit{Grade})_j + \gamma_{04}\,(\textit{Assessment Tools})_j + \gamma_{05}\,(\textit{Assessment Uses})_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

$$\beta_{2j} = \gamma_{20} + u_{2j}$$

$$\beta_{3j} = \gamma_{30} + \gamma_{31}\,(\textit{Feedback})_j + \gamma_{32}\,(\textit{Assessment Tools})_j + \gamma_{33}\,(\textit{Assessment Uses})_j + u_{3j}$$

$$\beta_{4j} = \gamma_{40} + \gamma_{41}\,(\textit{Feedback})_j + \gamma_{42}\,(\textit{Assessment Tools})_j + \gamma_{43}\,(\textit{Assessment Uses})_j + u_{4j}$$

$$\beta_{5j} = \gamma_{50} + u_{5j}$$

$$\beta_{6j} = \gamma_{60} + u_{6j}$$

This system of hierarchical linear equations allowed the modeling of achievement performance for student ***i*** in classroom ***j*** .

Subjects

The subjects of this study included the students and their mathematics teachers who fully participated in the Third International Mathematics and Science Study in the USA. In total, the database included 374 mathematics teachers and 10,973 students in grades seven (35%) and eight (65%). Of the students in the database, 4010 were deleted from subsequent analysis: 118 could not be linked to a teacher in the teacher database, 225 only completed one-half of the two-part assessment, and 3667 did not have teachers

with completed background questionnaires. Of the mathematics teachers, 46 were deleted because they did not have corresponding students in the student database (30), had fewer than six students in the student database (12), or had substantial missing data on questionnaire items of primary interest (4). The resulting database included 6963 students with 328 teachers who had completed background questionnaires and were in a class of at least six students with completed assessments and background questionnaires.

To briefly assess differences between the group of students in the final data set and those excluded, a simple mean difference in math score was evaluated, which resulted in a mean difference equal to about 0.16 standard deviations. The group excluded from further analyses as described above had, on average, mathematics performance scores about 0.16 of a standard deviation below the group included in the final data set.

Of the 6963 students, 51% were females and 49% were males; 36% were in 7^{th} grade and 64% were in 8^{th} grade (similar to the original sample). Nearly all of the students always spoke English at home (87.2%) while others sometimes spoke English at home (11.6%), and few never spoke English at home (1.2%).

Of the 328 mathematics teachers, 35 percent taught grade seven and 65 percent taught grade 8 (which matched the distribution of students). Just over 67 percent of the mathematics teachers were female while 33 percent were male. On average, teachers had 21 students in their class also included in the student database (s.d. = 6). Age and education levels of the teachers can be seen in Table 4.

Table 4
Age and Education Level of Mathematics Teachers

Age	Math Teachers
Under 25	5.2%
25-29	11.9%
30-39	21.0%
40-49	40.5%
50-59	18.0%
60 or more	2.1%
Level of Education	
BA, no teacher training	1.2%
BA, with teacher training	57.1%
MA, no teacher training	0.6%
MA, with teacher training	41.0%

Mathematics Assessment Instruments

The Third International Mathematics and Science Study (TIMSS) was conducted in 1995 with over 40 countries. It was the third in a series of studies conducted by the International Association for the Evaluation of Educational Achievement (IEA). It was the largest study of international educational achievement ever done. Two unique contributions of the TIMSS to previous studies included the in-depth analyses of curriculum, the addition of performance tasks, and the extensive teacher and student background questionnaires to evaluate the social and cultural contexts for learning. Data were collected from three populations (9-year-olds, 13-year-olds, and students in their final year of secondary education). Population 2 included the two grades with the highest proportion of 13-year-olds in each country, which usually included grades seven and eight. This project included population 2 from the United States of America (USA). Future studies could investigate these issues in populations 1 and 3 and with the

international data. For a more complete discussion of the TIMSS database, see Gonzales and Smith (1997).

Mathematics Assessment Frameworks. Three dimensions were used to define the curriculum frameworks used in TIMSS. Subject matter content included the content of the items; performance expectation described the performance or behavior a test item might elicit; and perspectives focused on students' attitudes, interests, and motivations. The mathematics framework as illustrated by TIMSS is presented in Figure 10.

Content and Performance Expectations were used to design the TIMSS assessment. There were six final content categories used for reporting purposes, including (1) fractions and number sense; (2) geometry; (3) algebra; (4) data representation, analysis, and probability; (5) measurement; and (6) proportionality.

Items were first grouped into clusters and the clusters were distributed throughout eight booklets so that each booklet had the same level of difficulty and content coverage. The core cluster was present in all eight booklets (appeared in the same position in each booklet), while some clusters (focus items) were included in three or four booklets, some clusters (breadth items) were included in only one booklet, some clusters (constructed-response items) were included in two booklets, where all were rotated for even distribution. The instrument was administered in two consecutive sittings, amounting to a total of 90 minutes of testing time.

Subsequently, each student was exposed to a single booklet that was equivalent in difficulty and content to other booklets, including both science and mathematics items in multiple-choice and constructed-response format.

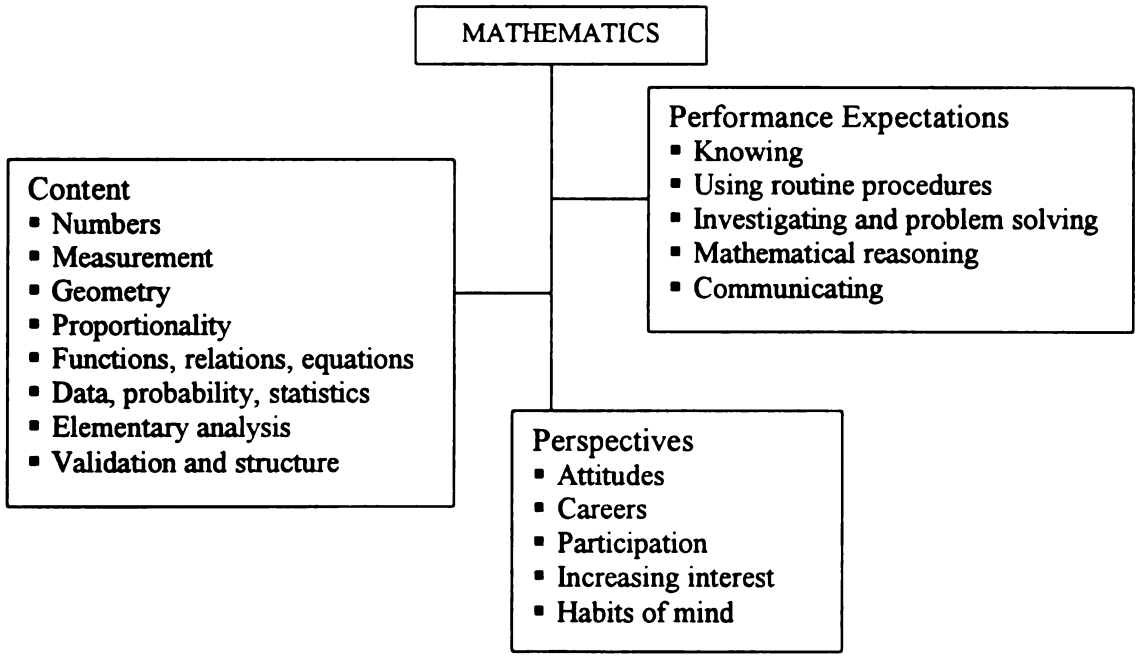

Figure 10. TIMSS Mathematics curriculum framework.

The international consensus building and item pool development took three years to complete. Each participating country had to agree that the test adequately accommodated their curriculum. The items themselves were written in three formats. The multiple-choice items included four or five options and students were directed to select the best answer. The constructed-response items included short answer or extended-response items and responses could include drawings as well as written answers. These items were scored using a one-, two-, or three-point rubric, depending on the complexity of the item. Items were scored in teams of 6 with team leaders. Approximately 10% of the responses were scored independently by two raters. The percent of exact agreement across all items was 99% for mathematics and 95% for science. The complete item pool consisted of 151 mathematics items (125 multiple-

choice, 26 constructed-response) and 135 science items (102 multiple-choice, 33 constructed response). The overall assessment reliabilities were reported as median Cronbach's alpha coefficients from the eight booklets: the mathematics assessment had a reliability of 0.86 among 7th grade students and 0.89 among 8th grade students.

Sampling and Sampling Weights. Sampling was done in two stages. First, schools were stratified based on region (Northeast, Southeast, Midwest, West), school type (public, private), and high and low levels of minority status (schools with high levels of minority status received twice the selection probability). Schools were sampled using a probability proportional to size stratified by the above characteristics. In all, 220 schools were sampled of which 179 participated at the 7th grade and 183 participated at the 8th grade.

Within each school, one 7th grade class and two 8th grade classrooms were sampled with equal probability. Since each student was essentially selected using probability sampling, the probability of each student being selected was known. The inverse of this probability was used as a sampling weight. The sum of the weights, in a properly selected and weighted sample, approximated the population size.

In seventh grade, 3886 of the 4168 students sampled were eligible (based on exclusion criteria for excluding special needs classrooms). The sum of the sampling weights was approximately 3,156,847 for seventh grade. In eighth grade, 7807 of the 7814 students sampled were eligible. The sum of the sampling weights was approximately 3,188,297 (for a total population of 7th and 8th grade students of 6,345,144). The total eligible sample, unweighted, was 11,693, of which 10,973 (94%) participated in the TIMSS assessment.

When conducting an analysis where population estimates are desired, the weighting of subjects will achieve proportionally represented estimates. However, two considerations are important here: the sampling procedure was based on (1) a systematic stratification of schools and (2) random selection of classrooms within schools.

The use of correcting weights was possible in most of the techniques employed in this study for estimating parameters of linear models. However, in cases where LISREL was used to conduct confirmatory factor analyses and latent variable structural equation modeling, it was not clear exactly how these weights were accounted. The statistics used in structural equation modeling included the covariance matrix of all variables in the model. The covariance matrix was computed with the appropriate weights. However, it was unknown whether or not the estimation of path coefficients and their standard errors were properly adjusted for the weights that were used to estimate the covariance matrix. Thus, the results of the LISREL analyses have been interpreted with caution.

Scoring and Scaling. The calibration of the TIMSS items for scoring students was completed using Quest Rasch software with maximum likelihood estimates of a scaled score, centering item difficulties at zero. They also considered items "not reached" as not administered in the item calibration, but considered the same items as incorrect in computing ability scores for each student.

The mathematics assessment was rescored for this project to address several concerns. The 225 students who only completed one-half of the assessment (missed one of the two administration sessions) were excluded from the calibration. A 2-parameter logistic model was used to account for differences in item discrimination (these differences are described below). Finally, all items assigned within a booklet were

considered administered for both item calibration and ability scoring. The original database did not include standard errors for the Rasch ability estimates, which would be needed to properly account for the heteroscedasticity in analyses using IRT ability estimates as outcomes. By rescoring the assessment, an estimate of standard error for each ability score was obtained.

The items were calibrated by using Multilog 6.3 (Thissen, 1991a, 1991b), which provided marginal maximum likelihood (MML) item parameter estimates where the latent variable was random. Once item parameters were estimated, they were then used to estimate maximum likelihood ability scores for each student. Since the results were simply used to evaluate the strength of certain relationships, the ability estimates were not rescaled, a step usually done to avoid reporting negative ability values. The items were centered at zero and the ability scores were based on the scale set by the item difficulty parameter.

Finally, a marginal reliability was provided which was an average reliability over levels of ability. This marginal reliability is an accurate representation of the precision of measurement when the test information is relatively uniform over ability levels.

TIMSS Background Questionnaires

The student questionnaires were administered separately from the assessment instruments. Completion of the questionnaires took between 20 and 40 minutes. The student questionnaires asked students about their demographics and home environment, including academic activities, living environment, parental education and expectations,

attitudes toward mathematics and science, and about their classroom experiences in mathematics and science.

The teacher questionnaires asked teachers about their own background, instructional practices, students' opportunity to learn and their pedagogic beliefs. There were separate questionnaires for teachers of mathematics and teachers of science.

In Appendix A, the items from the teacher and student background questionnaires are stated exactly as they appeared on the questionnaire. Tables of frequencies for each response to each question used in this project are also included.

Statistical Analyses

Preliminary data analysis included descriptive statistics of all data collected, including teacher and student perceptions, characteristics of teacher assessment practices, and student performance on the TIMSS assessment.

Structural Equation Modeling. A structural equation model was fit to the teacher-level portion of the data, as illustrated in Figure 5. This model was estimated using LISREL (Jöreskog & Sörbom, 1998).

The measurement model (including observed indictors) provided two kinds of information. The first kind included the factor loadings, the strength of the relationships between the observed indicator variables and a latent trait. The second set included the error variances associated with each indicator variable, the variance that was unique to the item and not accounted for by the latent trait. These error variances were then taken into consideration when estimating the relationships among the latent traits.

The structural model (including the latent traits) allowed estimation of the strength of relationships among the latent traits—the unobserved variables of interest. The SEM allowed for the simultaneous estimation of several equations or each path in the diagram. The measurement error in the latent traits is used to properly estimate the strength of the paths between latent traits.

LISREL has been commonly used to estimation structural equation models. It uses maximum likelihood estimation, a full information technique where all the parameters are simultaneously estimated. In addition, maximum likelihood estimators are known to be consistent and asymptotically efficient in large samples (Bollen, 1989). Finally, several model fit indices were provided and were described when presented.

Hierarchical Linear Modeling. A hierarchical linear model was fit to the data. The model as described earlier was fit to the data based on the sample as described. This model was estimated using HLM (Bryk, Raudenbush, & Congdon, 1996a, 1996b). Appropriate analyses were also conducted based on tests of coefficients and model modifications were made. HLM similarly tested for the significance in model-data fit between one or more models based on the inclusion or exclusion of certain estimated parameters.

HLM partitions variance and covariance components across levels, estimating the variance within classrooms and between classrooms. HLM provides improved estimation of effects within higher-level units, such as classrooms in this project. The student-level regression model was applied to each classroom, where the estimation of effects was improved through "borrowing strength from the fact that similar estimates exist for other classrooms" (Bryk & Raudenbush, 1992, p. 5). Empirical Bayes estimates

were computed for randomly varying student-level coefficients (i.e., coefficients that randomly varied across classrooms). Generalized least squares estimates were computed for classroom-level coefficients, employing the different levels of precision of information provided by each classroom. Finally, maximum likelihood estimates of the variance and covariance components at both levels were computed. HLM also enables the testing of effects that cross levels, including interactions between student and classroom characteristics (e.g., how teacher assessment practices might affect the relationship between self-efficacy and performance). For a more complete description of estimation in HLM, see Bryk and Raudenbush (pp. 32-56).

Additional multivariate techniques were used to confirm relationships prior to evaluating the full model, including the following.

Hierarchical Cluster Analysis. Hierarchical cluster analysis was used to identify clusters of class content topics and assessment tools and their uses, based on their similarity as employed by teachers. Using between-groups linkage and Pearson correlations for the clustering provides a result similar to a factor analysis. The clusters are parallel to factors in a factor analysis.

The strength of cluster analysis comes from its graphical display of the nesting or hierarchical clustering of variables (see Kachigan, 1991). For example, assessment tools were merged into clusters at different stages depending on their similarity as employed (weighted) by teachers. The distance scale is arbitrary, but illustrates the relative distance between each assessment tool as weighted by teachers. In the resulting graphical display, the distance needed to reduce the number of factors can be seen, starting on the left where each tool is a single factor, to the far right where all tools are combined into a common

factor. The distance between points where tools are clustered provides a relative indication of the information lost (or gained) by moving to the next level of clustering (where larger gaps indicate a larger loss of information as one moves from left to right--from more clusters to fewer clusters).

Discriminant Analysis. When the outcome or a characteristic of interest is categorical (or dichotomous), a useful method of analysis is discriminant analysis. Discriminant analysis was used to evaluate the degree to which assessment practices could be used to differentiate teachers with known characteristics, such as gender, frequency of assigning homework, or amount of algebra covered in a class. Discriminant analysis is parallel to regression analysis for categorical or qualitative outcomes (Kachigan, 1991). The discriminant analysis computes a weighted combination of variables (similar to a regression equation) to classify teachers into one of the known groups -- assigning each teacher a value on the categorical criterion variable.

One strength of discriminant analysis in this study was the resulting analysis of classification. The discriminant function was used to classify teachers based on a known grouping variable (e.g., gender) and their assessment practices. This discriminant classification given assessment practices can be compared to the actual group membership (e.g., since gender is known). In this way, the accuracy of classification can be used to evaluate the strength of the discriminant function. This procedure was used to identify teacher characteristics through which teachers could be differentiated based on their assessment practices. Another way of interpreting such results could be: do teacher assessment practices as a set, differ by gender (or level of education, type of class, etc.).

Levels of Inference

Both student and teacher (classroom) levels of information were used in this project. The HLM analyses were able to combine their simultaneous effects and test for effects that crossed the two levels as well. The appropriate level of inference from the TIMSS is the student within classroom. The classrooms and teachers are the focus of this project; however, they are considered the teachers of a representative sample of students, not a representative sample of classrooms. Although sampling occurred in several stages, the design weights for the TIMSS sample allow results to approximate a representative sample of middle school (13 year old) students. The unit of analysis, and appropriate level of inference, was student within classrooms.

CHAPTER IV

Results

These results are presented in order of the research questions posed earlier. However, before the research questions are addressed, a brief description of the classrooms included in these analyses and their average performance is presented.

Descriptions of Classrooms and Average Performance

In the full analysis, there were 328 mathematics classrooms. The unweighted average classroom TIMSS mathematics score was 0.02, with a standard deviation of 0.71, a minimum classroom average score of -1.47, and a maximum of 1.83. The distribution of classroom averages can be seen in Figure 11.

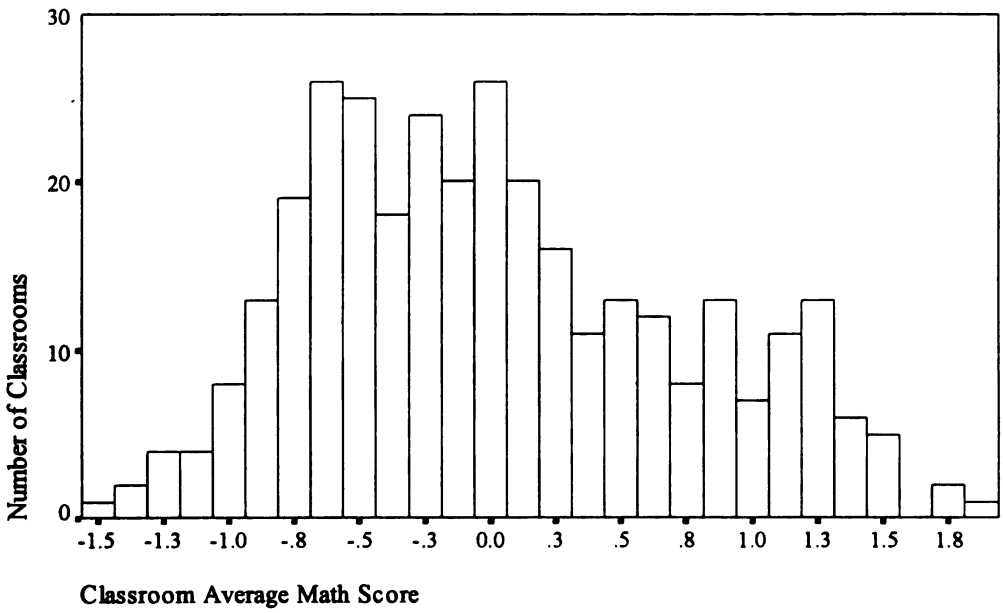

Figure 11. Distribution of classroom average mathematics scores.

The standard deviation of scores within classrooms was related to the magnitude of the average classroom score. As can be seen in Figure 12, the highest and lowest scoring classrooms had the lowest classroom standard deviation, they were more homogenous in performance than classrooms in the middle of the score distribution. The standard deviation of classroom performance was not related to class size; larger classrooms were just has heterogeneous (or homogenous) as were smaller classrooms.

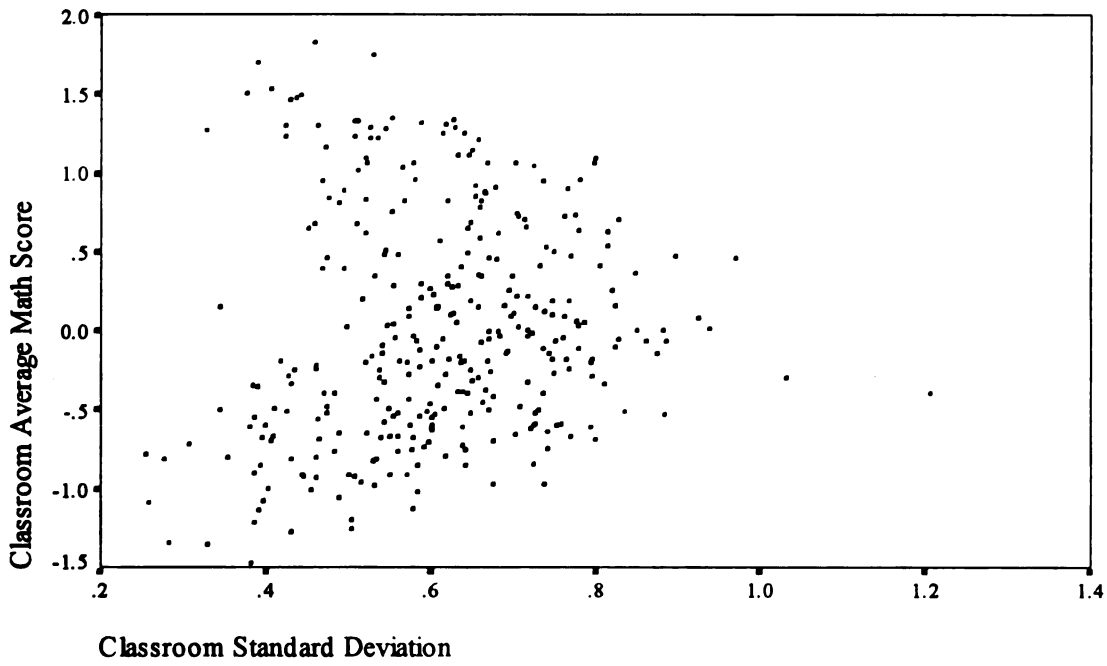

Figure 12. Scatterplot of average classroom score and classroom score standard deviation.

Course Content. TIMSS has characterized the U.S. mathematics curriculum as "a mile wide and an inch deep" (Schmidt, McKnight, & Raizen, 1996). This TIMSS conclusion came from analyses of textbooks and curriculum guides from the various

countries participating in TIMSS. There was no information available regarding the exact type of mathematics classes in which students were enrolled. From the database, however, information was available about the topics covered by teachers in their mathematics classrooms. Teachers were asked to rate how long they spent on each of 37 topics in their class during the year. The options included: not taught (0), taught in one to five periods (1), six to ten periods (2), 11 to 15 periods (3), or more than 15 periods (4).

Based on these ratings, a series of factor analyses were conducted to uncover the relationships among the 37 topics. These topics are listed in Table 6 (in the order they were presented in the questionnaire), with the mean level of coverage as reported by teachers (on a scale of 0-4 as described above). Slightly more descriptive definitions for each topic are reported in Appendix B, as they were presented on the questionnaire.

Based on brief descriptions and logical similarities among all topics, the topics were clustered into the six content areas of the mathematics assessment. These logical clusters were then compared to the results of factor analyses of the 37 topics using three different extraction methods: maximum likelihood, principal axis factoring, and principal components analysis; each time extracting factors with Eigen values over 1.0. Each result was slightly different; however, the similarities supported the original logical clustering with three modifications -- the inclusion of three different additional factors from the empirical results. The logical clustering and the empirical clustering results are summarized in Table 7.

The resulting nine factors were used in subsequent analyses when investigating topic coverage by mathematics teachers. The ratings of all topics within each factor were averaged to obtain an average topic coverage score.

Table 6

Mean Level of Coverage of Mathematics Topics by Teachers

Topics	*Mean*
A. Whole Numbers	1.69
Meaning of whole numbers; place value, numeration	1.14
Operations with and properties of whole numbers	1.56
B. Common & Decimal Fractions	2.02
Meaning, representation, uses of common fractions	1.35
Properties of common fractions	1.22
Meaning, representation, uses of decimal fractions	1.30
Properties of decimal fractions	1.17
Relationships between common & decimal fractions	1.22
Conversion of equivalent forms	1.27
Ordering of fractions	1.13
C. Percentages	1.79
D. Number Sets & Concepts	2.25
E. Number Theory	2.01
F. Estimation & Number Sense	1.61
G. Measurement Units & Processes	1.63
H. Estimation & Error of Measurement	0.70
I. Perimeter, Area, & Volume	1.76
J. Basics of One & Two Dimensional Geometry	1.52
K. Geometric Congruence & Similarity	0.89
L. Geometric Transformations & Symmetry	0.56
M. Construction & Three Dimensional Geometry	0.58
N. Ratio & Proportion	1.69
Concepts and meaning	1.05
Applications and uses	1.28
O. Proportionality: Slope, Trigonometry & Interpolation	0.27
Slope and trigonometry	0.30
Linear interpolation and extrapolation	0.10
P. Functions, Relations, & Patterns	0.75
Q. Equations, Inequalities, & Formulas	1.82
Linear equations and formulas	1.67
Other equations and formulas	0.85
R. Statistics & Data	1.02
S. Probability & Uncertainty	0.66
T. Sets & Logic	0.43
U. Problem Solving Strategies	1.98
V. Other Content: computers, nature of mathematics, proofs	0.79

Table 7
Topic Coverage Factors

Logical Topic Clusters	Factor Labels	Empirical--Final Factors
A B C D E F	Fractions & Number Sense	A B
I J K L M	Geometry	I J K L M
P Q	Algebra	D Q
R S	Probability & Statistics	R S
G H	Measurement	G H
N O	Proportionality	O P
	*Ratios, Proportions, Percentages	C N
	*Number Theory & Estimation	E F
	*Logic & Problem Solving	T U V

Note. The topic cluster letters correspond to topics as reported in Table 3.
* These three factors were not actual reporting categories for the TIMSS assessment. However, they were distinct factors in terms of topics covered by mathematics teachers.

In Table 8, the means and standard deviations are reported for the resulting topic coverage factors. Although number theory & estimation and algebra had the highest average level of coverage, as will be seen below, they were actually covered by teachers who did not spend so much time on other topics, and vice versa.

Table 8
Descriptive Statistics for Topic Coverage Factors

Variable	*Mean*	*SD*
Number Theory, Estimation	1.81	0.82
Algebra	1.59	1.03
Ratios, Proportions, Percentages	1.37	0.84
Fractions & Number Sense	1.26	0.74
Measurement	1.17	0.81
Logic, Problem Solving	1.07	0.80
Geometry	1.06	0.71
Probability & Statistics	0.84	0.88
Proportionality	0.38	0.56

To evaluate how these topics were taught together across teachers, a hierarchical cluster analysis was conducted (Figure 13). At the lowest level, geometry and measurement combined into a cluster or common factor. Following closely, algebra and proportionality combined into a cluster.

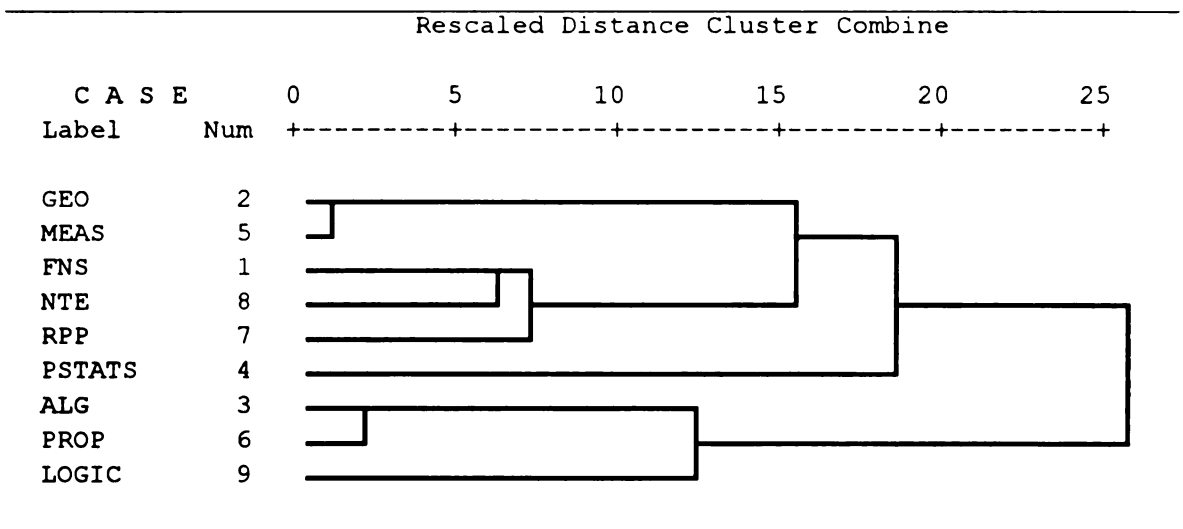

Figure 13. Display of hierarchical cluster analysis of topics covered.

Figure 13 illustrates the strong association between time spent teaching geometry and measurement in terms of topics as covered by teachers. That is, geometry and measurement clustered very early (about 2 on this scale); whereas these topics did not cluster with probability and statistics until much later (about 20 on this scale). Algebra and proportionality were also taught together to a high degree, even though proportionality was the one topic area covered the least by teachers, while algebra had a

high average level of coverage. However, this indicated that those teachers that did teach proportionality were also teaching algebra topics.

Fractions & number sense; number theory & estimation; and ratios, proportions and percentages were taught at similar levels among teachers. Probability and statistics, and logic and problem solving were two topic areas that were not associated with the other topics to a high degree as covered by teachers.

Do these topic factors make sense in terms of their relationship to overall math performance? Correlations between the mean level of topic coverage and the average math score (2-PL Thetas) for each teacher are reported in Table 9.

Table 9
Correlations for Topic Coverage Level with Mean Math Score

Topics Covered	Correlation with Math Score
Algebra	0.4364
Proportionality	0.3622
Logic, Problem Solving	0.1823
Geometry	0.0902
Probability, Statistics	-0.0307
Ratios, Percentages	-0.0614
Measurement	-0.1225
Number Theory, Estimation	-0.2074
Fractions, Number Sense	-0.3588

Based on Table 9, the overall math score was most highly related to the level of coverage of algebra and proportionality topics. Several of the other areas with midrange coverage made little to no difference on the overall math score. Coverage of fractions & number sense, however, had a negative impact on overall math score. This was likely

because teachers who covered fractions & number sense were not covering algebra and proportionality, which appeared to be highly related to performance. This suggested negative correlations between time spent on algebra and proportionality with fractions. This was in fact what was obtained (see Table 10).

Table 10
Intercorrelations of Average Time Spent on Topic Factors

	FNS	GEO	ALG	PSTAT	MEAS	PROP	RPP	NTE
Geometry	.15							
Algebra	-.20	.10						
Probability, Statistics	.10	.29	.13					
Measurement	.27	.53	-.03	.20				
Proportionality	-.18	.19	.51	.13	.06			
Ratios, P Percentages	.43	.22	.06	.27	.29	-.02		
Number Theory, Est	.44	.20	.04	.15	.41	-.01	.38	
Logic, P Solving	-.10	.25	.36	.26	.27	.28	.04	.14

Table 10 contains intercorrelations between average time spent on each topic factor. The correlation between average time spent on algebra and proportionality was among the highest of any pair of topics ($r = 0.51$). The highest correlation was between measurement and geometry ($r = 0.53$), which was also evident from the cluster analysis results. As expected, the correlation of fractions & number sense with algebra was negative ($r = -0.20$).

Although there was no direct information available about the exact nature of the mathematics classes taken by students, four types of classes were likely: (1) remedial mathematics, (2) regular 8th grade mathematics, (3) pre-algebra, and (4) algebra. A

problem which TIMSS presented previously was that especially in the case of regular 8th grade math courses, there was a wide range of topics covered at all different levels of rigor, thus, the characterization of a "splintered" curriculum (see Schmidt, McKnight, & Raizen, 1996). Based on these results, two indicators seemed to differentiate the performance level of the classrooms, (1) algebra and (2) fractions and number sense (hereafter referred to as fractions). The difference between time spent on algebra and fractions illustrates the degree to which these were taught at similar levels, as displayed in Figure 14. This difference covered the range of -4.0 (i.e., maximum coverage of fractions and no time spent on algebra) to +4.0 (i.e., no coverage of fractions and maximum time spent on algebra). The time spent on algebra versus fractions was related to classroom average mathematics scores, $r = 0.52$.

To provide a parsimonious indicator of relative prior math achievement of students within classrooms, the difference in time spent on algebra versus fractions was used in analyses of the full model. This difference was referred to in this study as "high-low relative prior math achievement" or simply "relative prior achievement." This was an important issue since the argument did not include the role of curriculum, per se, but the importance of considering the role of prior achievement or pre-requisite skill levels of students.

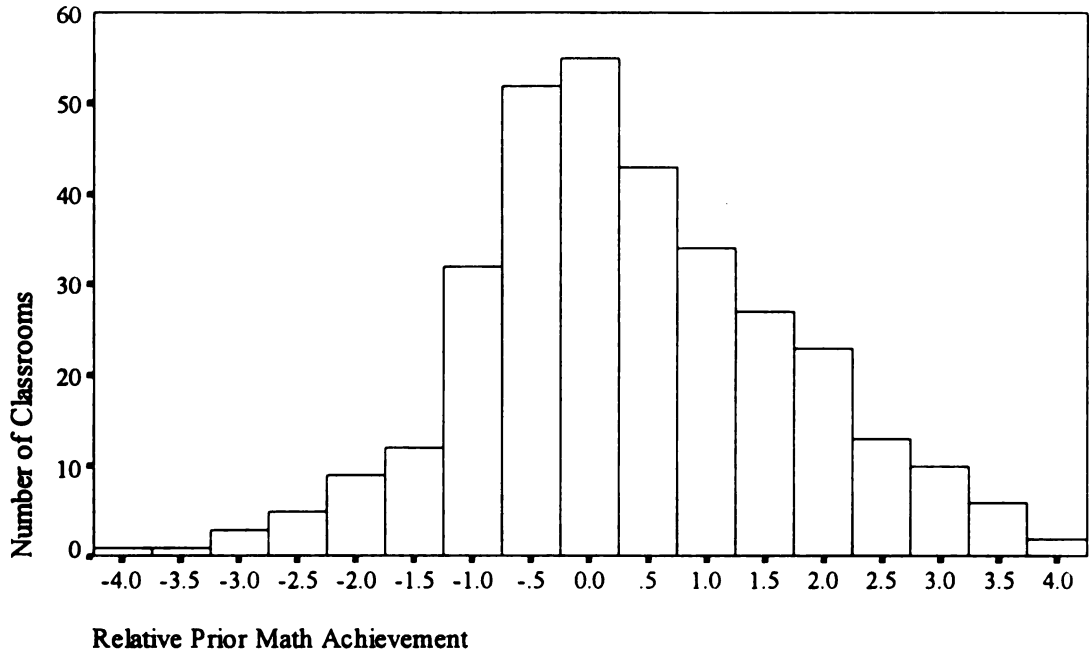

Figure 14. Number of classrooms by relative prior math achievement (the time spent on algebra versus fractions).

Course Content & Grade. There was a significant difference in the amount of time spent on algebra topics in seventh- and eighth-grade classrooms. Twice as much time was spent on algebra topics in eighth-grade classrooms than in seventh-grade classrooms. The amount of time spent on algebra topics was likely a strong indicator of the type of math class students were in, which may be strongly related to several other variables that were impossible to differentiate (i.e., prerequisite skills of students, rigor of instruction, and others which should differ by grade). Figure 15 illustrates the distribution of time spent on algebra topics by grade. Approximately three percent of the seventh-grade classrooms spent more than four weeks on each of three algebra topic areas while 18 percent of the eighth-grade classrooms did so.

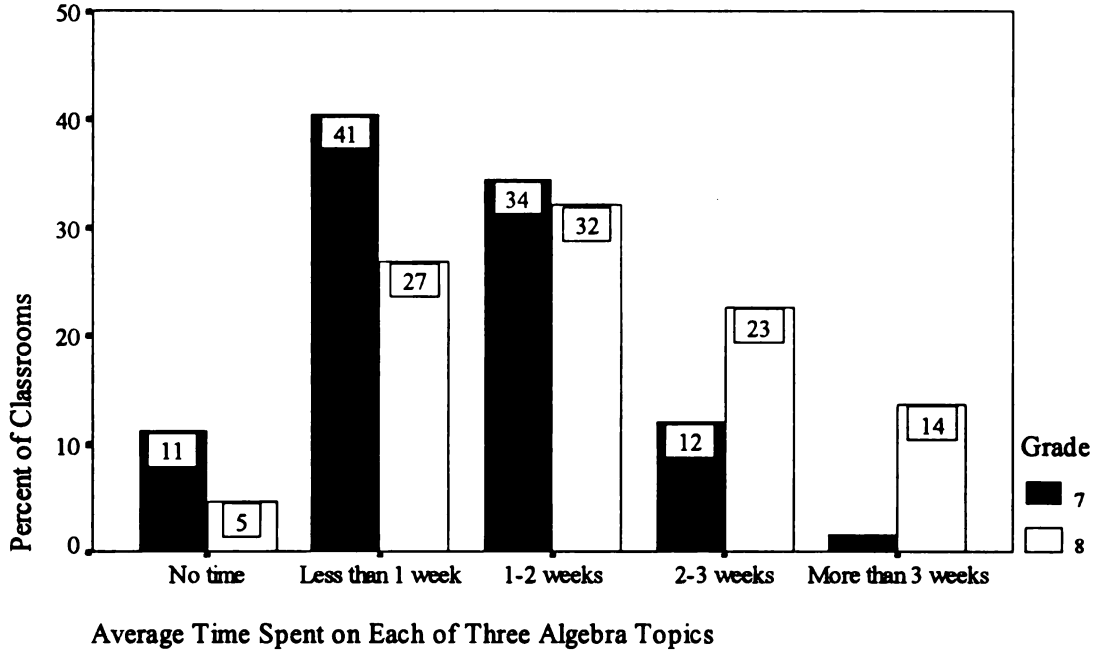

Figure 16. Time spent on algebra topics by grade.

Through ninth and 12th grade, two percent of all students completed credits in functional math, 11 percent in basic math, and 43 percent in pre-formal math, which included pre-algebra (Davenport, Davison, Kuang, Ding, Kim, & Kwak, 1998). Nearly half of all middle school students were not likely to reach pre-algebra before entering high school.

Course Content & Gender. The average proportion of females in each classroom was 0.51. There was one all-male seventh-grade classroom with 25 students. This classroom did not cover any algebra topics. There were also five classrooms that were more than 80 percent male, ranging in size from six to 12 students (as included in the database). There was one all-female eighth-grade classroom with 11 students. This classroom also did not cover any algebra topics. In addition, there were four classrooms that were more than 80 percent female, ranging in size from nine to 28 students.

In Figure 17, a scatterplot displays the proportion of females by the average math score for each of the 328 classrooms. The correlation between these two variables was $r = 0.06$, not significantly different from zero.

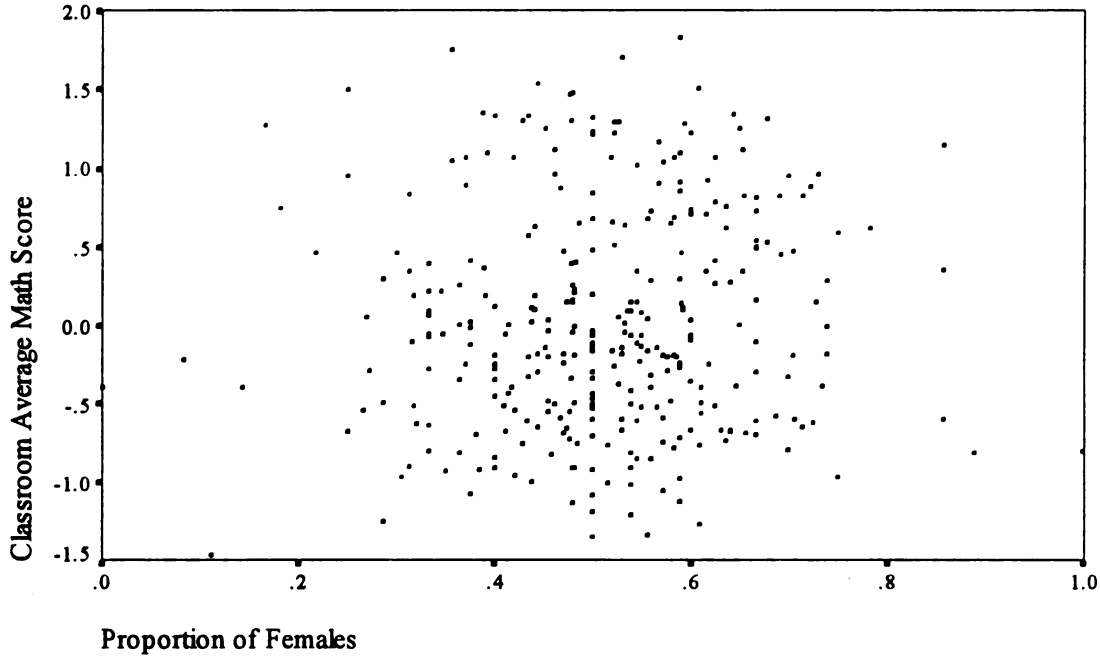

Figure 17. Scatterplot of classroom math score and proportion of females.

There was essentially no difference in the proportion of females in a class and the amount of time spent on algebra topics. This suggests that there was no reason to believe that females were systematically left out of algebra or pre-algebra classes. The correlation between the full algebra coverage variable and proportion of females in each class was $r = 0.03$, not significantly different from zero. By the time students reach high school (9^{th} grade), they are equally prepared for advanced mathematics, however, fewer females will complete credits in the advanced sequences in high school (Davenport et al.,

1998). Even so, Davenport et al. reported that males complete more functional, basic, and preformal courses (sequences prior to algebra) as well as advanced courses than females, a trend that paralleles the larger variance in mathematics achievement.

Course Content & English. The average proportion of students who always or almost always spoke English at home was 0.88. There was one classroom with 13 students where less than 8 percent of the students spoke English at home. This classroom had covered each of the three algebra topic areas on average for one to five periods. There were also four other classrooms with less than 40 percent of students who spoke English at home, ranging in size from 14 to 25. All of these classrooms had covered each of the three algebra topic areas less than 11 periods and included a mix of seventh and eighth-grade classrooms.

There were 91 classrooms where 100 percent of the students spoke English at home. The algebra coverage in these classrooms included the full possible range. Figure 18 displays a scatter plot of the proportion of students who spoke English at home and the average math score for each of the 328 classrooms. The correlation between these two variables was $r = 0.33$, significant and moderate, but also potentially nonlinear. All of the primarily non-home-English speaking classrooms (from 0.0 to 0.4) scored below the classroom average math score (0.02).

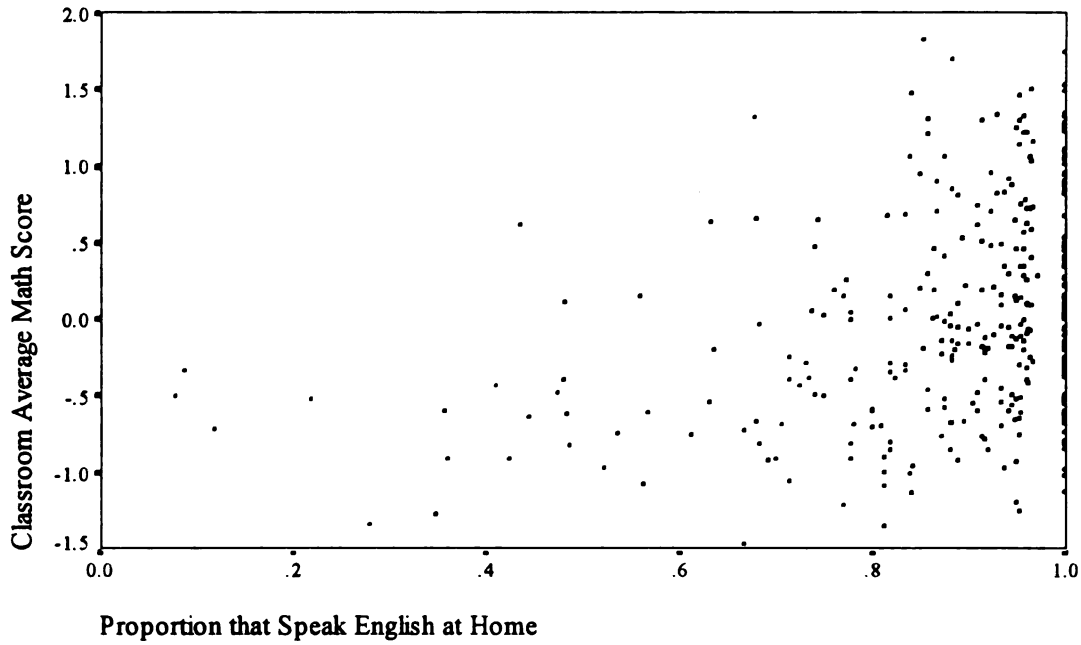

Figure 18. Scatterplot of classroom math score and proportion who speak English at home.

There was a similar relationship between proportion that spoke English at home and the amount of time spent on algebra topics in the class. Figure 19 illustrates this relationship, using a combined range of time spent on algebra for simplicity of display. Classrooms that spent more time on algebra topics were primarily composed of students who spoke English at home (including more than 70% of the students in a classroom). The correlation between the level of algebra coverage and proportion of students who spoke English at home was $r = 0.19$, significant but small; and again this was possibly a nonlinear relationship.

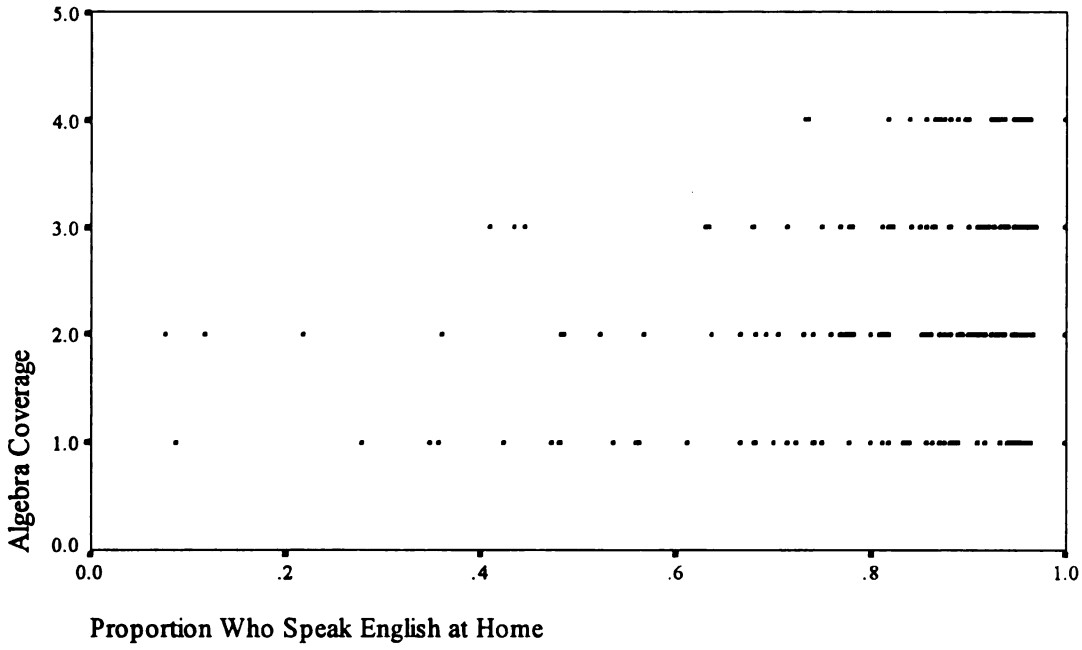

Figure 19. Scatterplot of time spent on algebra topics and proportion that speak English at home for each classroom.

Table 11 provides the means for several of these variables by time spent on algebra topics. The correlations reported previously with respect to algebra coverage become clearer from this table, where no relationship existed with proportion of females or size of the class, and a positive relationship existed with proportion who speak English at home and overall classroom average math score.

Table 11

Mean values for classroom characteristics by time spent on algebra topics

Time Spent on Algebra Topics	Proportion Female	Proportion who Speak English at Home	Classroom Average Math Score	Average Classroom Size
0.00	.44	.78	-.59	23
0.33	.50	.84	-.50	23
0.67	.54	.85	-.17	19
1.00	.50	.88	.02	21
1.33	.52	.87	-.18	20
1.67	.51	.90	.14	21
2.00	.51	.87	.11	22
2.33	.52	.91	.33	22
2.67	.52	.91	.30	23
3.00	.51	.92	.25	19
3.33	.57	.90	.41	20
3.67	.50	.99	.60	19
4.00	.48	.92	.76	24

Note. Time spent on algebra topics was an average of time given three topic areas described earlier, on a range of 0 (no time) to 4 (more than 15 periods). Over the three specific algebra topics, this included 0 to over 45 periods.

What Are the Current Assessment Practices of Teachers

Homework Assignments. Nearly all of the mathematics teachers in the sample (99%) reported that they assigned homework; three of the 328 teachers did not assign homework. Most teachers assigned homework three or four times a week (57%) or every day (27%). According to teachers, the average homework assignment usually took 15-30 minutes (70%) or 31-60 minutes (19%) to complete.

The types of homework tasks included in the survey of mathematics teachers included: working problem/question sets in textbooks, completing worksheets or workbook tasks, finding one or more uses of the content covered, small investigations of gathering data, reading in a textbook or supplementary materials, working individually on long term projects or experiments, writing definitions or other short writing assignments, working as a small group on long term projects or experiments, preparing oral reports either individually or as a small group, and keeping a journal.

The 99 percent confidence intervals for the frequency means are displayed in Figure 20, based on the scale of never (1) to always (4). Mathematics teachers most often assigned textbook problems and worksheets or workbook assignments. They least often assigned oral reports or journal writing.

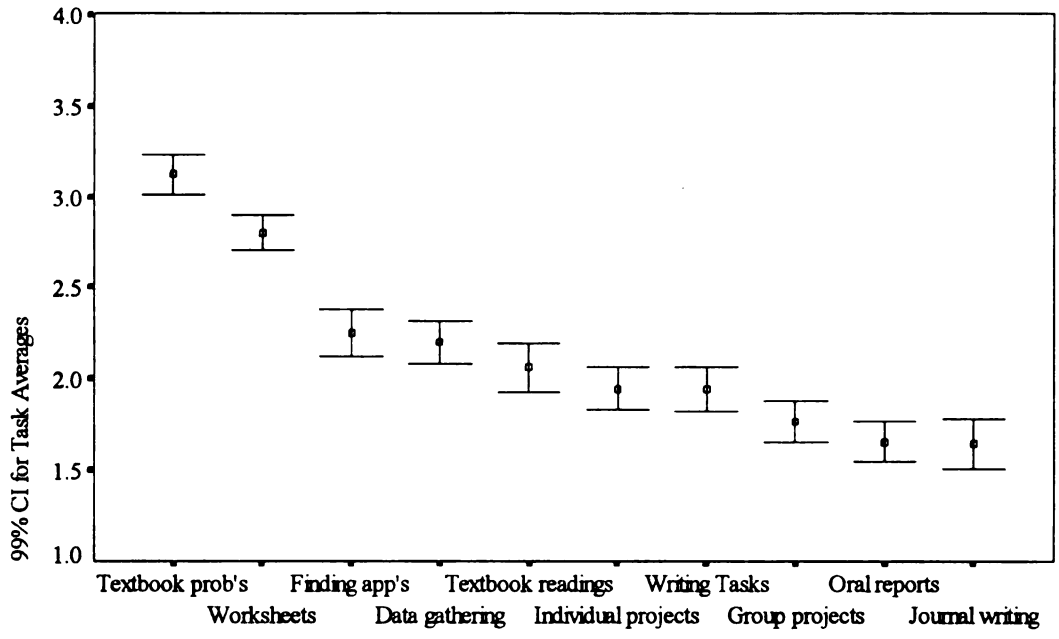

Figure 20. Error bars displaying the 99% confidence interval for frequency of assignment of homework tasks.

A cluster analysis was completed on the frequency of assignment of homework tasks (see Figure 21). Gathering data, working individually, and working in small groups clustered early and were highly related. Oral reports and finding applications of course content clustered together at the next two stages. Writing, reading, and journal work were also less likely to occur with the previous types of homework. Finally, worksheets or workbook tasks and problem sets in textbooks were not tightly associated and not likely to be given by teachers who assigned other types of homework tasks. These relationships are also evident from their intercorrelations reported in Table 12.

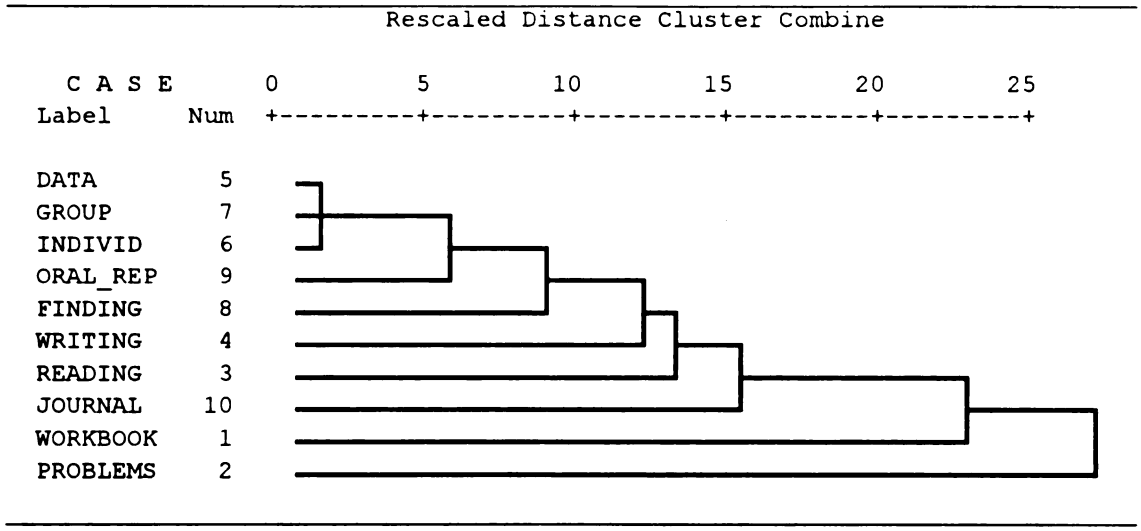

Figure 21. A cluster analysis of homework assignment tasks.

Table 12
Intercorrelations of Homework Assignment Tasks

	Workbk	Prob's	Read	Write	Data	Ind	Group	Find	Oral
Text Problems	-.31								
Text Reading	.06	.16							
Writing	.02	.03	.32						
Data Collection	.05	-.12	.39	.41					
Individual Project	.01	-.15	.21	.24	.60				
Group Project	.16	-.17	.29	.29	.62	.61			
Finding Uses	.02	.01	.38	.35	.52	.39	.44		
Oral Report	.03	-.08	.31	.31	.46	.47	.53	.34	
Journal Writing	.11	.04	.19	.31	.22	.24	.30	.25	.29

In addition, correlations between frequency of homework tasks assigned and the indicator of relative prior math achievement were low. The strongest positive correlations indicated that classes with higher relative prior math achievement had teachers that were more likely to assign textbook problems ($r = 0.18$) than worksheets

(r = -0.16). These teachers included those that taught algebra more often compared to fractions (high-level versus low-level mathematics classrooms).

Another dimension of homework included the uses or purposes of homework -- what teachers did with the homework once it had been completed. Among the teacher survey response options, teachers could record whether or not the homework was completed; use it to contribute toward students' grades; give feedback on homework to the whole class; have students correct their own assignments in class; use it as a basis for class discussion; collect, correct assignments, and return to students; have students exchange assignments and correct them in class; or collect, correct, and keep the assignments.

The 99 percent confidence intervals of the means for uses of homework are displayed in Figure 22, on the scale of never (1) to always (4). Three groupings were evident. Recording completion, contributing to grades, and providing feedback to the class were the most common uses for homework. Having student correct each other's assignments and keeping assignments were the least common -- although teachers did these things between rarely and sometimes.

A cluster analysis of the frequency of uses for homework suggested that using homework to give feedback to the whole class and as a basis for class discussion were highly related in terms of frequency of use. Also, recording completion of assignments and using homework as a basis for grades were closely related. Other uses of homework were weakly associated and less common. These relationships are displayed in the cluster analysis in Figure 23.

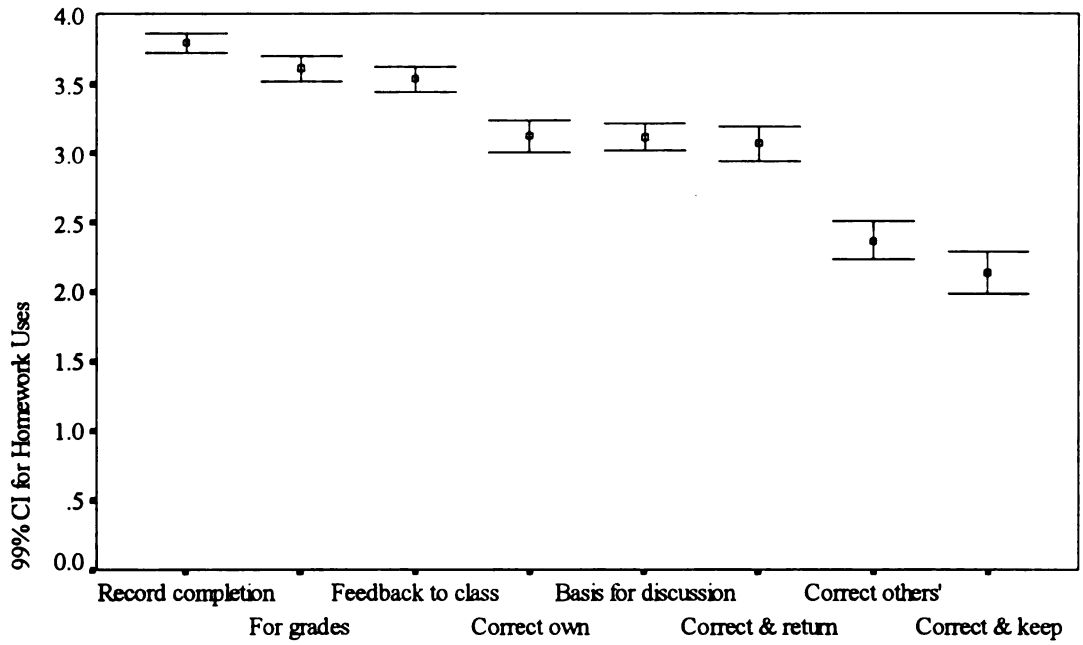

Figure 22. Error bars displaying the 99% confidence interval for average uses of homework.

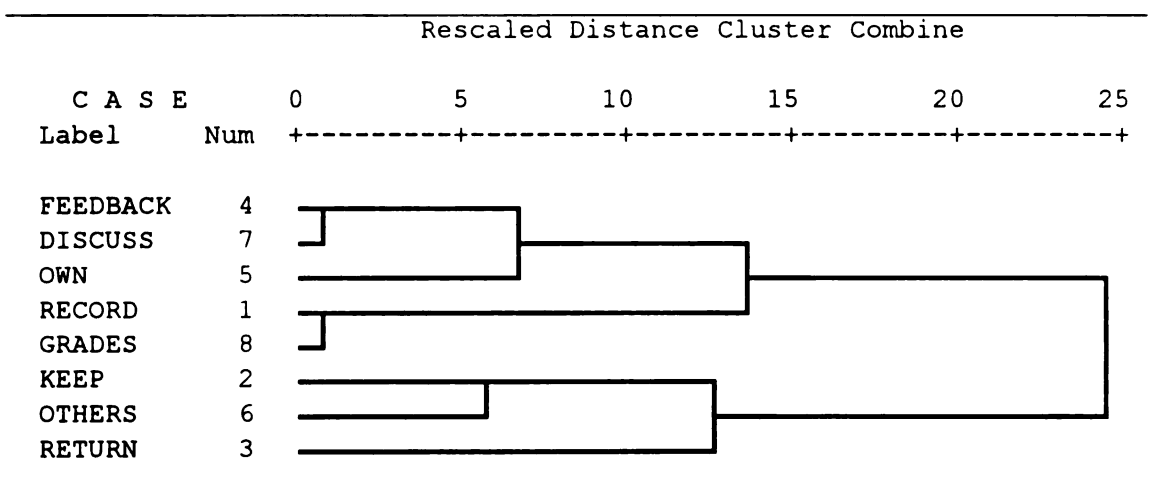

Figure 23. A cluster analysis of uses of completed homework assignments.

Again, the relationships in the cluster analysis were based on intercorrelations, similar to a factor analysis, as can be seen in the table of correlations (Table 13). Also included in Table 13 are the intercorrelations between types of homework and homework uses. Few of these correlations were 0.20 or greater. Textbook reading tasks and individual and group projects were assigned more often in classrooms where teachers used homework tasks for discussion and feedback. Short writing tasks were assigned more often in classrooms where teachers were more likely to collect and keep homework assignments.

Table 13
Intercorrelations for Homework Assignment Uses and Tasks

	Record	Keep	Return	Feedback	Own	Others	Discuss	Grades
Keep	-.04							
Return	.21	.15						
Feedback	.21	-.06	.03					
Correct Own	.09	-.12	-.28	.23				
Correct Others	-.04	.27	.16	.01	-.17			
Discuss	.10	.00	-.11	.36	.28	-.07		
Provide Grades	.35	-.07	.08	.23	.07	-.12	.11	
Worksheets	.09	.02	.07	-.00	.05	.06	.03	.04
Textbook Prob's	-.07	-.09	-.08	.09	.07	-.02	.08	.05
Reading Tasks	-.01	.11	.05	.21	.11	.11	.20	-.04
Writing Tasks	.09	.20	.18	.06	.02	.19	.06	-.03
Data Gathering	.16	.11	.16	.16	.12	.11	.12	.11
Individual Proj's	.15	.08	.13	.18	.09	.12	.20	.07
Group Project	.17	.17	.11	.20	.07	.15	.15	.08
Finding Uses	.11	.03	.11	.16	.07	.16	.14	.03
Oral Report	.06	.10	.05	.13	.04	.02	.14	.02
Journal Writing	-.03	.07	.06	.08	-.01	.12	.12	-.05

Classroom Assessment Tools. A second facet of classroom assessment practice included the various tools used by mathematics teachers to assess the work of their students. Survey response options for teacher assessments included teacher-made short answer or essay tests that required students to describe or explain their reasoning (subjective); scores from homework assignments; observations of students; responses of students in class; scores from projects or practical exercises; teacher-made multiple-choice, true-false, and matching tests (objective); and standardized tests produced outside the school. Teachers rated the amount of weight they gave each type of assessment as they assessed the work of their students. Note, however, that although one of the distinguishing characteristics between what were termed here as teacher-made objective and subjective tests was the requirement that students "explain their reasoning," it should be recognized that teacher-made objective tests could require reasoning skills as well; although students did not have to "explain their reasoning" on these items.

The 99 percent confidence intervals of the mean weights for the various types of assessments are illustrated in Figure 24, on the scale of none (1) to a great deal (4). Teacher-made subjective tests and homework were weighted the most by teachers, while teacher-made objective tests and tests produced outside the classroom were weighted the least; however, differences were small overall.

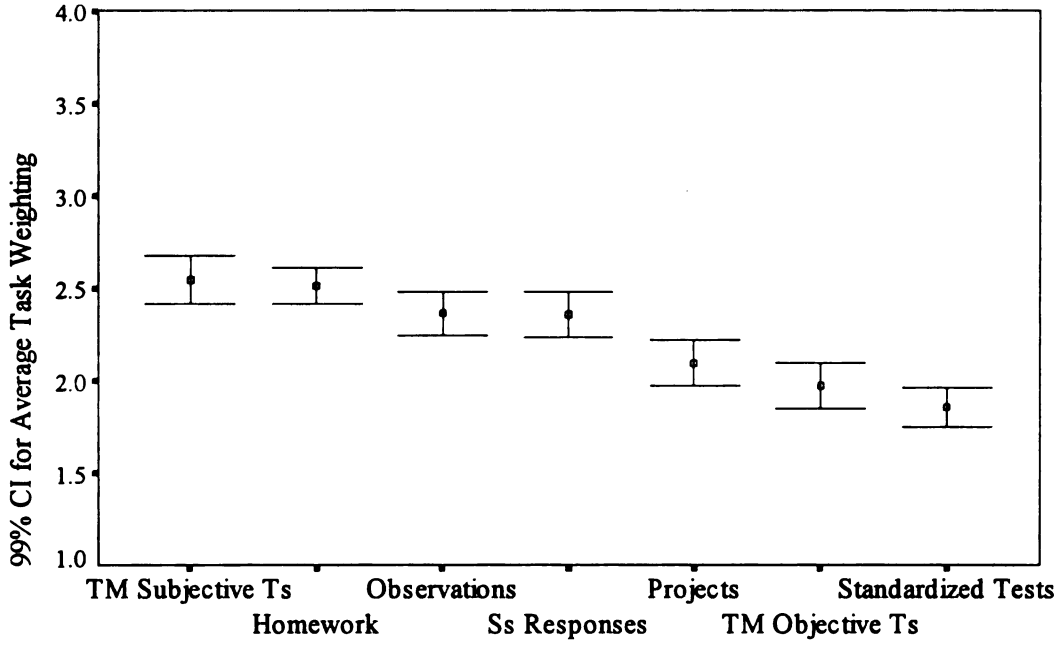

Figure 24. Error bars displaying the 99% confidence interval for average weights for types of assessments employed by teachers.

A cluster analysis of the types of assessments used by teachers (see Figure 25) suggested that observation and responses were weighted similarly; they were closely clustered early on. All other types of assessments were clustered later, but relatively at the same point. The use of standardized tests produced outside the school was also weakly associated with the other tools. This suggested that one factor includes observations and student responses, loosely related to other types of more objective tasks, and another factor included the use of standardized measures. These relationships can also be seen in the correlation matrix, Table 14.

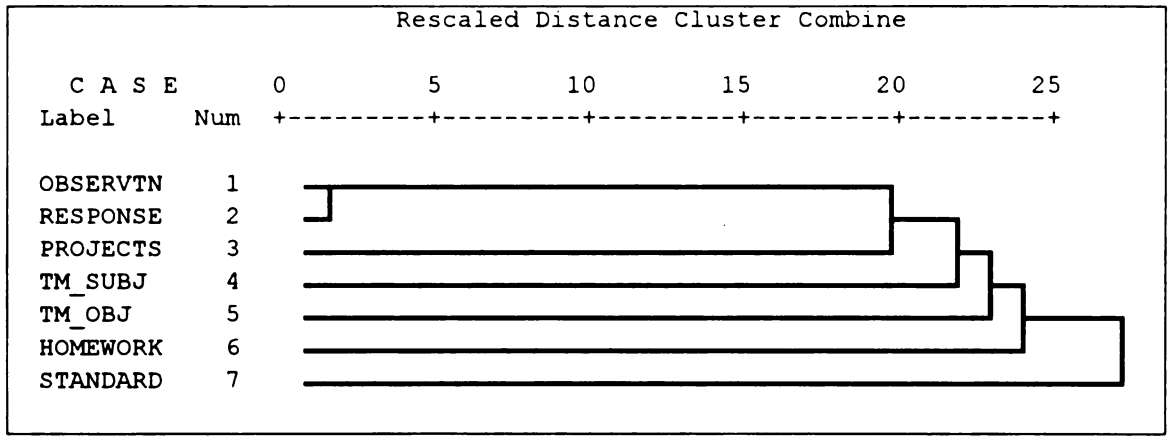

Figure 25. A cluster analysis of assessment tools.

Table 14
Intercorrelations of Assessment Tools

	Teacher-Made Tests					
	Subjective	Objective	Homework	Projects	Observ'n	Responses
TM Objective Tests	.22					
Homework	.10	.17				
Projects	.28	.23	.16			
Observations	.25	.21	.24	.31		
Student Responses	.26	.24	.27	.29	.79	
Standardized Tests	.01	.17	.07	-.05	.17	.22

In the classroom practice facet of assessment tools was a second dimension describing the uses of those tools. Teachers could use assessment information they gathered from students to provide students' grades, provide feedback to students, plan for future lessons, report to parents, diagnose students' learning problems, or assign students to different programs or tracks.

The 99 percent confidence intervals for the mean uses are displayed in Figure 26, on the scale of none (1) to a great deal (4). Most often, teachers used assessment

information to provide for students' grades and feedback to students. Teachers were least likely to use assessment information to assign students to different programs or tracks, although teachers sometimes did this.

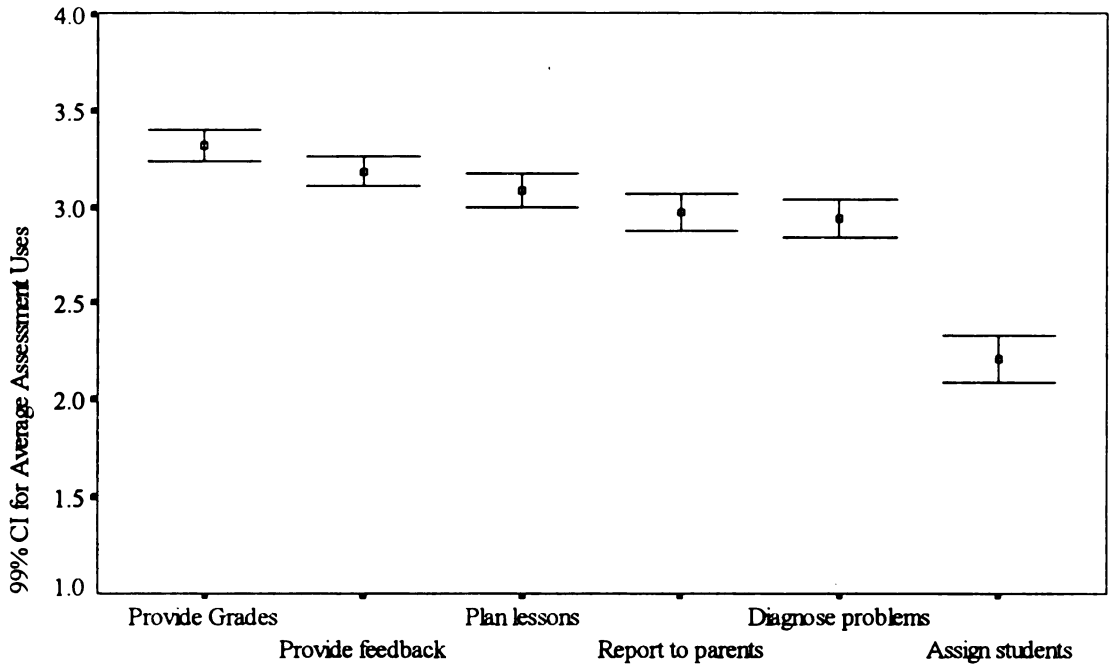

Figure 26. Error bars displaying the 99% confidence interval for average uses of assessment information.

A cluster analysis (Figure 27) of the uses of assessment information suggested that grading and feedback to students were tightly associated; that is, teachers who used assessment information for grading students were also likely to use it to provide feedback to students. Diagnosing learning problems and planning for future lessons were also associated, but weakly connected to other uses. This also seemed likely since diagnosis of learning problems was also informative for instructional feedback; uncovering what

areas students are weak in provides useful information for additional instruction. Finally, the assignment of students to special programs or tracks was not associated with the other uses and teachers were least likely to employ this use of assessment information.

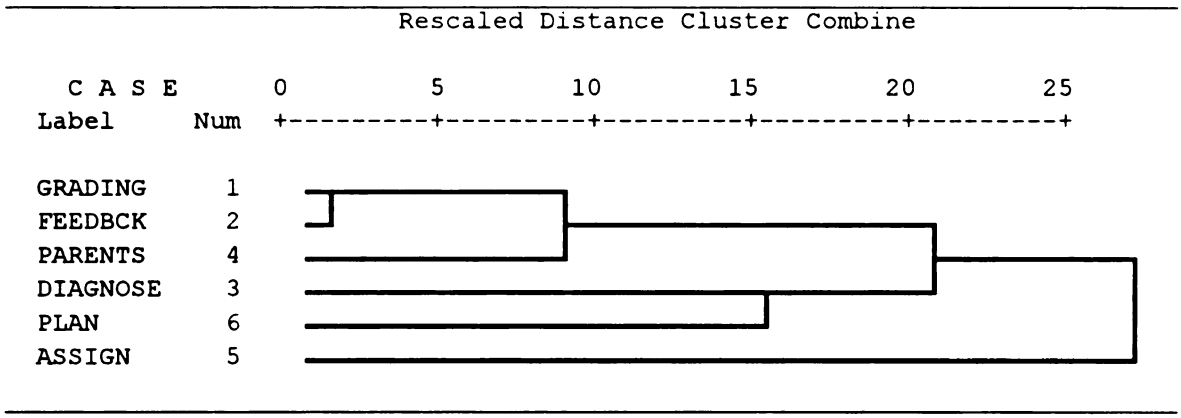

Figure 27. A cluster analysis of uses of assessment information.

The correlation table (Table 15) of uses for assessment information shows such a pattern. Finally, Table 15 also reports the intercorrelations between assessment tasks and uses of assessment information. Again, most correlations were small. The strongest relationships were among teachers who weight heavily teacher-made subjective tests and use assessment information for grading, feedback, and diagnosing learning problems. Also, there was a strong relationship between teachers who give more weight to observations and responses of students and those who use assessment information for diagnosing problems and planning future lessons.

Table 15
Intercorrelations of Uses for Assessment Information

	Grading	Feedback	Parents	Diagnosis	Plan	Assign
Feedback	.55					
Parent Reports	.46	.49				
Diagnosis	.31	.51	.44			
Plan Lessons	.24	.39	.31	.41		
Assign Groups	.22	.21	.39	.36	.31	
Standardized Tests	.03	.02	.08	.08	.05	.07
T-M Subjective Tests	.20	.29	.17	.20	.14	.11
TM Objective Tests	.00	.08	.15	.14	.09	.11
Homework	-.09	-.01	.03	.05	.04	.11
Project Performance	-.00	.17	.18	.16	.16	.08
Observation	-.04	.09	.08	.22	.19	.10
Student Responses	-.05	.10	.04	.20	.21	.14

Summarizing Classroom Assessment Practices. Two facets of classroom assessment practices were identified and investigated with multiple indicators. These included homework tasks and other assessment tools. Both of these facets were described in two dimensions, including the tools or tasks that could be used and the uses or purposes for each.

In the facet of homework, teachers seemed to work with three types of assignments: (1) workbook or worksheet tasks, (2) textbook problems, and (3) individual or group projects and data gathering tasks. Teachers most commonly either (1) recorded completion of the homework assignments and used them to assign students grades or (2) used completed assignments to provide feedback to the whole class or as a basis for class discussion. Few of the intercorrelations between the two dimensions of types of homework and uses of homework were greater than 0.20, which suggested independence. In addition, none of the types or uses of homework were correlated with class-type or

relative prior math achievement indicators (i.e., all were between -0.16 and 0.18). The strongest relationship with relative prior achievement occurred between the frequent use of worksheets in classes with greater fractions content and the frequent use of textbook problem sets in classes with greater algebra content.

In the facet of other assessment tools used by mathematics teachers, they employed (1) assessments that they had constructed themselves, including teacher-made subjective tests, objective tests (less so), and projects; or (2) general observations of students during class and student responses during class. They frequently either used the assessment information they gathered to (1) provide for students grades and to provide students with feedback or (2) for planning purposes both in terms of planning future lessons or diagnosing student learning problems. Again, the intercorrelations between tools and uses were small, generally 0.20 or less, which suggested a fair amount of independence between these dimensions. In addition, all of the relationships between tools and uses with relative prior math achievement indicators were small (i.e., all were between -0.13 and 0.04. The strongest relationship was between relative prior achievement and higher weighting of teacher-made objective tests to assess students in class (r = -0.13).

To summarize these findings, a series of confirmatory factor analyses were conducted using LISREL 8.20 (Jöreskog & Sörbom, 1998) to examine the interrelationships of these constructs relating to classroom assessment practices. A single confirmatory factor analysis was conducted on both the types of assessment tools used and the uses of assessment information. The results are illustrated in Figure 28, which displays three types of estimates: (1) uniquenesses or the variance due to unique factors

not included in the model, (2) factor loadings for each observable indicator (in rectangles) given the common latent variable or factor (in ovals), and (3) the correlation between the latent variables--factors. These parameters are in order as presented in the figure from left to right. All of the factor loadings were standardized (correlations between observed and latent variables) and significant.

This model was a good fit to the data. Common fit indices include a chi-square (χ^2) test statistic to assess the discrepancy between the correlation matrix implied by the model and the original observed correlation matrix. The resulting $\chi^2 = 23.9$, $df = 21$, $p = 0.30$. This suggested that the model fit well. Another common statistic is the root mean squared error of approximation (RMSEA), based on the analysis of residuals. The RMSEA = 0.021, where values less than 0.05 indicate excellent fit.

The interesting component was the intercorrelations of the latent factors across the two dimensions, tools and uses. The intercorrelation matrix of latent factors is presented in Table 16. The interesting correlations were those between the types of assessment tools used and the uses of assessment information. These four correlations ranged between 0.10 and 0.45. The strongest correlation was using student work for planning purposes ($r = 0.45$, $p < .05$), whereas the weakest and only non-significant correlation was using observational tools for providing feedback to students ($r = 0.10$, *ns*).

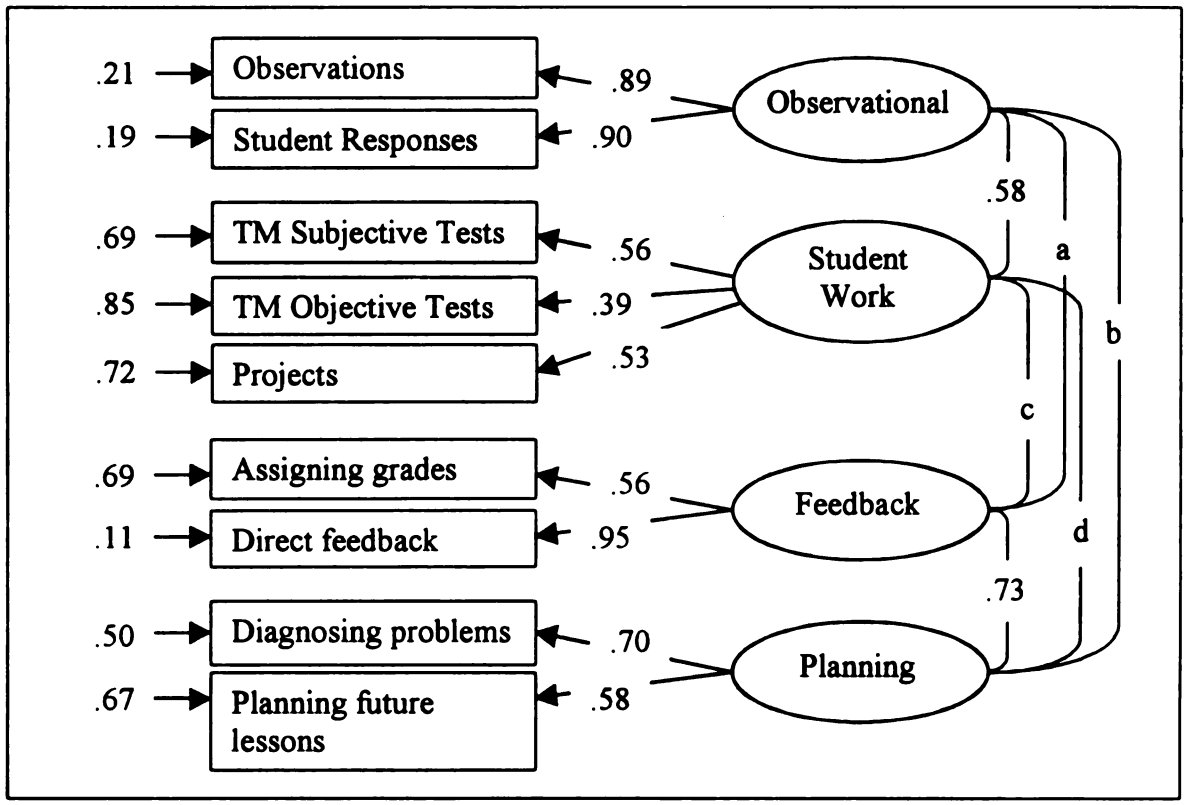

Figure 28. Structural equation model of primary assessment tools and uses.
Note. Values for latent variable intercorrelations (a-d) are reported in Table 16.

Table 16
Latent Factor Intercorrelations

	Types of Tools	
	Observational	Student Work
Uses of Information		
Feedback	0.10 (a)	0.39 (c)
Planning	0.35 (b)	0.45 (d)

Student-level Constructs

The primary goal of this work was to uncover the relationships between teacher classroom assessment practices and student performance on standardized assessments. However, there were three mediating constructs at the student-level that were presented earlier, based on theoretical and empirical grounds. These included the nature of the assessment feedback students received, student self-efficacy in the subject matter, and student effort in the subject matter.

The Nature of Assessment Feedback. Unfortunately, no direct questions were asked of teachers regarding the quality or kinds of feedback they provided to students based on their evaluation of assessment information. However, there were indicators of the kinds of feedback given to students from the students' own reports. Each one was a weak substitute for more direct questions.

Three indicators regarding possible kinds of feedback students received regarding the results of the assessments they completed included (1) correcting the homework of other students, (2) having the teacher correct homework, and (3) discussing completed homework as a class. Each of these activities provided some feedback to students regarding their performance on assessment tasks, primarily homework.

An initial description of the three indicators included descriptive statistics (Table 17) and intercorrelations (Table 18). The mean was based on a scale of never (1) to almost always (4). The frequencies of responses for Teacher checking homework and discussing homework were negatively skewed, since most students reported doing these two things more than "pretty often."

Table 17

Descriptive Statistics for Kinds of Feedback Students Receive on Homework

Variable	*Mean*	*SD*	*n*
Teacher checks homework	3.30	0.94	6811
Discuss completed homework	3.26	0.97	6817
Students check each other's homework	2.36	1.19	6802

Table 18

Correlations for Kinds of Feedback Students Receive on Homework

	Teacher Checks	Check Each Other's
Students check each other's homework	-0.13	
Discuss completed homework	0.24	0.08

Based on these statistics, it appeared that most students had opportunities to discuss their homework as a class, homework that was usually checked by the teacher. If teachers checked students' homework, they were more likely to discuss the homework as a class than if the students checked each other's work. Students who reported that they discussed homework in class also had higher overall math scores, as displayed in Figure 29. The difference between the *never* and *almost always* group means was about 0.28 standard deviations--a small effect size.

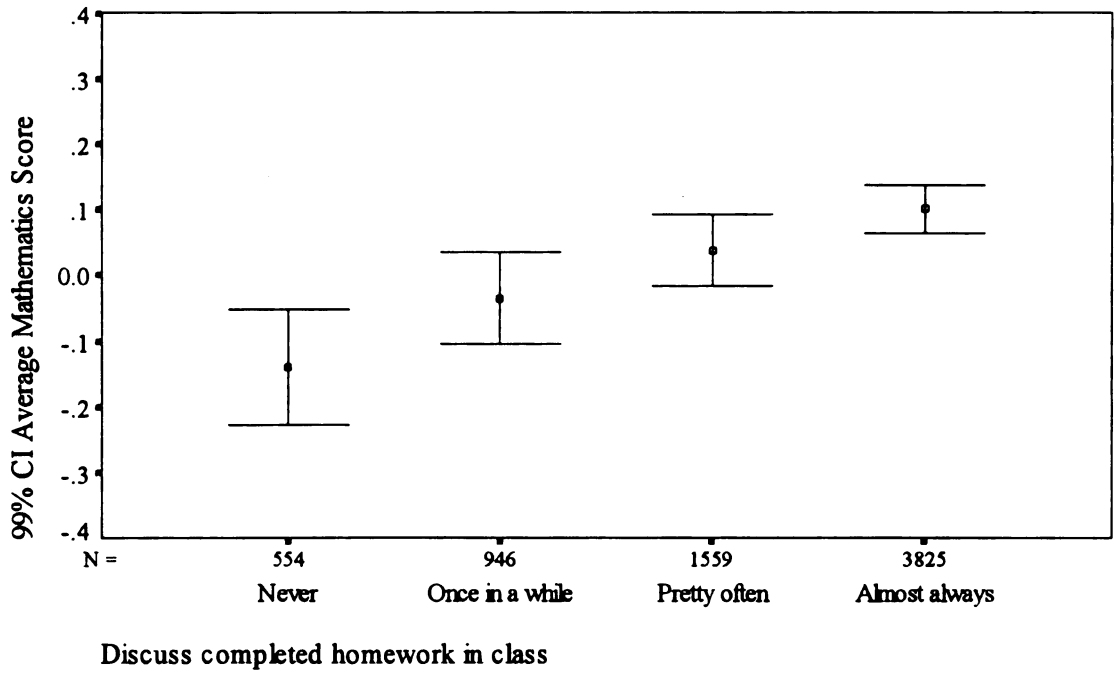

Figure 29. Confidence intervals (99%) for mean mathematics score by frequency with which completed homework was discussed in class, as reported by students.

These reports from students were in fair agreement with teacher reports. Students who reported that their teachers frequently checked homework had teachers who reported to record completion of homework (r = 0.24), although it was not a strong agreement. Students who reported that they checked each other's homework were in agreement with teacher reports about students correcting each other's work (r = .56). Student who reported that the class discussed completed homework were not in strong agreement with their teachers reported frequency of using homework for class discussion (r = 0.22). There was a great deal of variance in student reports about these activities within classrooms.

Student Self-Efficacy. Several items in the TIMSS background questionnaire addressed student academic self-concept regarding mathematics, student attribution of

control in mathematics and student perceptions of their potential for mastery of mathematics. These issues could be construed as a measure of self-efficacy in the subject matter. Among indicators of academic self-concept was an item asking students if they agreed with the statement: "I usually do well in mathematics." Regarding attribution of control, students reported the degree to which they agreed that to do well in mathematics at school, they need (1) lots of natural talent, (2) good luck, (3) lots of hard work studying at home, and (4) to memorize the textbook or notes. As possible indicators of potential for mastery, students reported how much they liked mathematics, and whether they enjoyed learning mathematics, thought math was an easy subject, and would have liked a job that involved using mathematics.

Table 19
Descriptive Statistics for Student Self-Efficacy Indicators

Variable	*Mean*	*SD*	*n*
Usually do well in math	3.21	0.73	6828
To do well in mathematics:			
Need natural talent	2.58	0.82	6818
Need good luck	2.27	0.88	6814
Need lots of hard work studying	3.40	0.73	6824
Need to memorize notes	2.68	0.90	6808
Like mathematics	2.86	0.92	6834
Enjoy learning math	2.86	0.83	6818
Math is an easy subject	2.48	0.92	6777
Would like a job involving math	2.43	0.97	6778

Generally, student responded to these items positively (Table 19). Students agreed that they usually did well in math. They liked math and enjoyed learning math, but were neutral regarding math being an easy subject or liking a job involving math. They agreed that to do well in mathematics, they needed to study hard but only slightly agreed regarding the importance of memorizing notes. They were neutral about needing talent to do well in math and slightly disagreed that they needed good luck.

Table 20
Inter-correlations of Self-Efficacy Indicators

	Talent	Luck	Study	Notes	Like	Enjoy	Easy	Job
Need good luck	.47							
Need lots of hard work studying	.05	-.03						
Need to memorize notes	.23	.23	.37					
Like mathematics	-.01	-.15	.12	.00				
Enjoy learning math	.06	-.09	.17	.08	.72			
Math is an easy subject	.04	-.04	-.04	-.02	.44	.42		
Would like a job involving math	.08	-.06	.12	.08	.50	.54	.35	
Usually do well in math	-.01	-.17	.05	-.07	.51	.48	.48	.35

From the correlation table (Table 20), the expected relationships appeared to hold true. Needing talent and good luck were moderately correlated ($r = 0.47$); these were both "uncontrollable" attributions. Needing to study hard and memorize notes were also moderately correlated ($r = 0.37$); these were both "controllable" attributions. Comparatively, the other correlations in this set of items were much smaller.

Students who reported to usually do well in mathematics were more likely to like math, enjoyed learning math, thought math was an easy subject, and would have liked a job involving math. In addition, the attribution items (talent, luck, study, memorize) were weakly correlated or uncorrelated with the other perceptions of mathematics.

There was no gender difference in self-efficacy; however, males were slightly more likely to make more uncontrollable attributions. There was also no gender difference in students' reports about usually doing well in mathematics.

Student Effort. The effort students put forth was a complex characteristic. Effort could have been displayed in many ways. From the TIMSS student background questionnaire, several indicators of student effort were available, which could also have been construed as indicators of motivation. Students reported the amount of time they spent studying math after school, the degree to which they thought it was important to do well in mathematics, whether they took notes in class and whether they started their homework in class. Oddly, the correlations among these items were very small (Table 21). Part of the problem with these correlations was that nearly all of the students agreed that it was important to do well in math (97%), while 67 percent and 77 percent reported that they took notes and began their homework in class. The other problem was with the way the questions were stated. The questions about taking notes and beginning homework were asked in a very general way: "How often do these things happen in your mathematics lesson?" They were not specifically asking students how often they themselves engaged in these activities. These were poor indicators of student effort.

Table 21
Intercorrelations for Potential Indicators of Student Effort

	Important to do well in math	Copy Notes
Copy notes from the board	.09	
Begin homework in class	.08	.04

Given the weak correlations among these indicators of student effort, a single indicator was used. The amount of time students spent studying math after school on a normal school day was the only indicator with a reasonable amount of variance. About 17 percent reported to spend no time studying math while 57 percent spent less than one hour, 24 percent spend one to two hours, and two percent spent three or more hours.

In Figure 30, the relationship between four attitudinal items and time spent doing homework, another interpretation became evident. Students who spent no time on average studying math agreed less that they did well in math and enjoyed learning math less than students who spent less than one hour or one to two hours on average. They also agreed more that math was boring more than any other group. Students who studied more than five hours on average were the least likely to agree that they usually did well in math. The time spent studying math appeared to be an indicator of both level of skill and attitude toward mathematics, and was likely related to effort among students who study less than three hours on average.

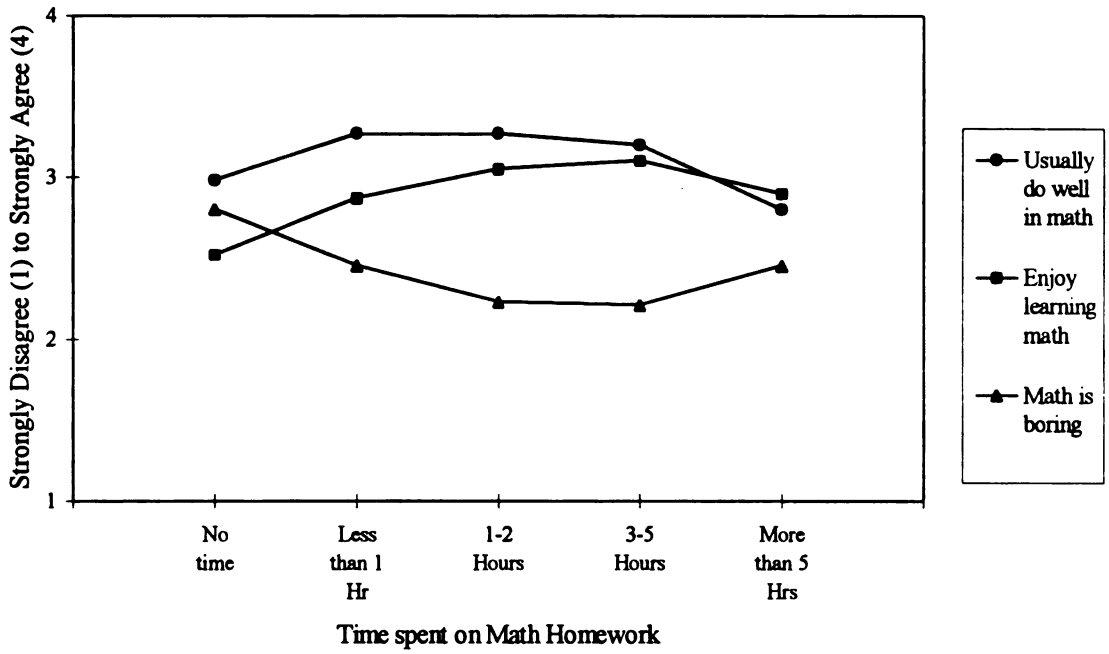

Figure 30. Attitudes toward mathematics by time spent on math homework.

Relationships between Teacher Practices and Classroom-level Achievement

Analyses at the classroom-level were based on teacher characteristics and practices. The average achievement level of the classroom was based on the number of students in the class who completed the TIMSS assessment. The number of students in each classroom ranged between 6 and 37, with an average of 21 students per class. The unweighted average classroom performance was 0.02 with a standard deviation of 0.71, as described earlier in the description of classrooms and performance (see Figure 11).

When estimating the average performance of students within a classroom, information on the standard error of the IRT ability parameter was lost. A piece of the information regarding the number of students in each classroom and the precision of their ability estimate (its standard error) was retained by using the hierarchical linear modeling (HLM) estimation procedures.

HLM allowed for the modeling of individual effects at one level and classroom effects at a second level (as described earlier), as well as the testing of interactions between levels. Weighted least squares estimates were computed and the standard error of ability estimates were employed, as was the students' probability of selection. By doing so, the heteroscedasticity of errors inherent in IRT parameters was addressed directly, as was the sampling of students with known probability. Finally, the variance was partitioned between classrooms and within classrooms. At this level of analysis, the primary concern was in accounting for variance in achievement between classrooms.

Classroom-Level Achievement. At the first stage, an unconditional HLM model was specified and estimated. This was similar to a one-way analysis of variance with

random effects, where the classroom means were considered random. This was an effective way to partition variance in the outcome within classrooms and between classrooms. The unconditional model was one without any explanatory variables. The unconditional model for mathematics achievement performance of students within classrooms was:

$$\text{Achievement}_{ij} = \beta_{0j} + r_{ij} \quad (7)$$

$$\beta_{0j} = \gamma_{00} + u_{0j} \ , \quad (8)$$

where the achievement score for student *i* in classroom *j* was a function of the classroom mean (β_{0j}) and the deviation of student *i*'s score from their classroom *j*'s mean score (r_{ij}).

Classroom means (β_{0j}) can then be modeled as a function of the overall grand mean (γ_{00}) and the deviation of classroom *j*'s mean from the grand mean (u_{0j}). The fixed effect was the overall grand mean (γ_{00}). The random effects were the deviations of the students from their classroom mean (r_{ij}, unique student effects) and the classroom means from the grand mean (u_{0j}, unique classroom effects). These random effects were assumed to have a mean of zero and a constant variance (all assumptions are evaluated below).

The variance of the student deviations from their classroom mean was the within-classroom variance (σ^2). The variance of the classroom deviations from the grand mean was the between classroom variance (τ). The estimates of the fixed and random effects are displayed in Table 22.

Table 22

Unconditional HLM Model of Student Mathematics Achievement Performance

Fixed Effects		*Coefficient*	*S. Error*	*T-Ratio*	*p-value*
Intercept L2, grand mean	γ_{00}	0.007	0.037	0.203	0.839

Random Effects		*Variance Component*	*df*	*Chi-Sq*	*p-value*
u_{0j}	τ_{00}	0.4214	327	7570	0.000
r_{ij}	σ^2	0.3630			

The standard table of results contains the estimated fixed effects (i.e., parameter estimates) and the random effects (i.e., variance components) at each level. Table 22 provides several pieces of information.

Based on the unconditional HLM model, using weighted least squares (weighting for the standard error of IRT ability estimate and students' probability of selection), the maximum likelihood point estimate of the grand mean (γ_{00}) mathematics achievement score was 0.007, essentially zero ($t = 0.20$, $p = 0.84$), slightly lower than the unweighted mean of 0.02. The standard deviation of average classroom achievement was 0.65, slightly less than the unweighted standard deviation of 0.71. The variance of classroom deviations (u_{0j}) from the grand mean was significantly different than zero ($\tau_{00} = 0.4214$, $\chi^2 = 7570$, $df = 327$, $p < 0.001$). Classrooms accounted for about 54 percent of the variance in students' mathematics achievement performance; the intraclass correlation was 0.54. This suggested that classrooms differed significantly in terms of their average performance. Subsequent models used classroom-level performance (i.e., β_{0j}) as the outcome to explain the between classroom variance, τ_{00}.

Generally, studies of academic achievement using HLM have found about 10 to 33 percent of the variance due to schools (Bryk & Raudenbush, 1992). However, in this study, classrooms are the organizational unit. It was reasonable to expect classroom-level achievement to vary to a higher degree than school-level achievement. School means should vary less than classroom means to the extent that schools include a larger population and greater diversity in student ability as compared to a classroom, particularly in a system where students enroll in classes based on prerequisite knowledge and skill or where a high degree of tracking occurs.

Finally, an estimator of the reliability (λ_j) of the sample mean in each classroom ($\bar{Y}_{.j}$) was also computed from the estimated variance components, where

$$\hat{\lambda}_j = \text{Reliabilit y}\,(\bar{Y}_{\bullet j}) = \hat{\tau}_{00} / \left[\hat{\tau}_{00} + \left(\hat{\sigma}^2 / n_j\right)\right].$$

Generally, the reliability of the sample classroom mean ($\bar{Y}_{\bullet j}$) as an estimate of the true classroom mean varies as a function of the number of students in each classroom. An overall measure of the reliability is the average of the classroom reliabilities, $\hat{\lambda} = \sum \hat{\lambda}_j / J$. The average reliability of classroom means was 0.95, which indicated that the sample means were highly reliable as indicators of the true classroom means. (See Bryk & Raudenbush, 1992, p. 63, for a more complete discussion.)

Relative Prior Math Achievement. At the classroom-level, the correlation between relative prior math achievement (based on indicators described above) and the average achievement level of each class was evaluated earlier ($r = 0.52$).

In an evaluation of the impact of prior achievement on overall mathematics score, an HLM analysis revealed the significance of the effects of relative prior math

achievement accounting for between classroom variance. A model including the high-low relative prior math achievement indicator was

$$\text{Achievement}_{ij} = \beta_{0j} + r_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\,(\text{Prior Math Achievement})_j + u_{0j},$$

where γ_{01} was the effect of the prior achievement indicator and u_{0j} was the deviation of classroom *j* from the grand mean of performance scores, controlling for classroom-level relative prior achievement. The results of the analysis are reported in Table 23.

Table 23
HLM Model of Student Mathematics Achievement Given Prior Achievement

Fixed Effects		*Coefficient*	*S. Error*	*T-Ratio*	*p-value*
Intercept Level-2, grand mean	γ_{00}	0.009	0.032	0.289	0.772
Prior Achievement	γ_{01}	0.249	0.023	10.921	0.000

Random Effects		*Variance Component*	*df*	*Chi-Sq*	*p-value*
u_{0j}	τ_{00}	0.3020	326	5579	0.000
r_{ij}	σ^2	0.3630			

Basically, classrooms with higher relative prior math achievement had higher average performance scores, as expected. This indicator provided an important control for prior achievement based on pre-requisite skill of student based on the type of mathematics (content and rigor) taught in a given classroom, and concomitantly the type of students enrolled in the class. Prior achievement explained just over 28% of the between-classroom variance (after all other variables described below were added to the model, prior achievement explained 13% of the between-classroom variance).

Homework Assignments and Uses. The relationship between homework assignments and achievement, including both the types of assignments and the uses of completed assignments, can be viewed in terms of average classroom performance. Again, there were no significant groupings of types of assignments, as described above, except for the combination of individual projects, group projects, and data gathering projects. The correlations between the homework assignment variables assessed earlier and average classroom achievement are reported in Table 24.

Table 24
Correlations between Types of Homework Assigned and Achievement Scores

Types	*r*	Types	*r*
Textbook problems	.23	Short writing assignments	.01
Textbook reading	.13	Data gathering projects	-.01
Individual projects	.07	Group projects	-.02
Oral reports	.03	Finding uses for lessons	-.02
Journal writing	.02	Workbook/worksheets	-.18

The frequency of homework assigned had a small but significant relationship with achievement ($r = 0.25$, not shown). Few of the types of homework assignments given in mathematics classes had a strong relationship with mathematics achievement scores. Of the types of homework assigned, the use of textbook problems had the strongest positive relationship, although small, with total scores ($r = 0.23$), while the use of workbooks or worksheet assignments had the largest negative relationship with total scores ($r = -0.18$).

Once the frequency of homework as assigned by the teacher was accounted for, the type of homework assigned made little to no difference in terms of explaining

variation in classroom achievement scores. Teachers who assigned homework more frequently were also the teachers who assigned textbook problems more frequently ($r = 0.29$). The assignment of textbook problems may have had a slight positive association with achievement while the assignment of worksheets had a slight negative association with achievement. The estimates of effects in this model are reported in Table 25. This model explained 32 percent of the original variance between classrooms.

Table 25
HLM Model of Student Mathematics Achievement Given Prior Achievement & Homework Frequency

Fixed Effects		*Coefficient*	*S. Error*	*T-Ratio*	*p-value*
Intercept Level-2, grand mean	γ_{00}	0.011	0.031	0.350	0.726
Prior Achievement	γ_{01}	0.238	0.022	10.606	0.000
Homework frequency	γ_{02}	0.164	0.040	4.068	0.000

Random Effects		*Variance Component*	*df*	*Chi-Sq*	*p-value*
u_{0j}	τ_{00}	0.2858	324	5203	0.000
r_{ij}	σ^2	0.3637			

The second piece of information regarding homework assignments was the purpose of the assignments as used by teachers. Based on earlier discussions, there were at least two distinguishable uses of homework assignments, including uses for individual feedback and class discussion as well as recording completion and grading. These two broader uses could be phrased as "direct feedback" to students and "evaluation-grading" by teachers.

Table 26

Correlations between Uses of Homework Assignments and Achievement Scores

Uses	*r*	Uses	*r*
Feedback to students	.12	Contribute to grades	-.06
Class discussion	.06	Record completion	-.08.

Feedback to students had the strongest positive correlation with achievement scores (although weak; see Table 26). The other uses of assignments had very small correlations, all less than 0.10, with achievement. When recording homework completion and grading were combined into a common variable (evaluation-grading) and feedback to students and class discussion were combined (direct feedback), only the evaluation-grading variable had an effect on achievement. The effect of feedback dissipated when evaluation and grading were included in the model. The final HLM results are reported in Table 27.

Classrooms where teachers more frequently used homework assignments for administrative uses had lower average achievement levels than classrooms otherwise. The inclusion of evaluation-grading of homework assignments had explained an additional three percent of the remaining variance between classrooms -- not a large amount, but enough to include in the model. This model explained nearly 34 percent of the original between-classroom variance.

Table 27

HLM Model of Student Mathematics Achievement Given Prior Achievement, Homework Frequency, and Uses of Homework Assignments

Fixed Effects		*Coefficient*	*S. Error*	*T-Ratio*	*p-value*
Intercept Level-2, grand mean	γ_{00}	0.008	0.031	0.271	0.786
Prior Achievement	γ_{01}	0.240	0.022	10.796	0.000
Homework frequency	γ_{02}	0.197	0.041	4.749	0.000
Evaluation-Grading	γ_{03}	-0.185	0.060	-3.097	0.002

Random Effects		*Variance Component*	*df*	*Chi-Sq*	*p-value*
u_{0j}	τ_{00}	0.2782	322	5027	0.000
r_{ij}	σ^2	0.3648			

Assessment Tools and Uses. Similar to homework assignments, the second facet of teachers' assessment practices included other assessment tools and their uses. As discussed above, there were two primary types of assessment tools mathematics teachers employed, tests and observations of students. To facilitate the evaluation of the effect of the employment of these tools on achievement, their correlations with achievement scores are reported in Table 28.

Table 28

Correlations between Assessment Tools and Achievement Scores

Tools	*r*	Tools	*r*
Teacher-made subjective tests	.04		
Projects	-.01	Observations of students	-.06
Responses of students	-.05	Teacher-made objective tests	-.16

Once again, none of the correlations were large in magnitude. The largest correlation was between teacher-made objective tests and achievement (r = -0.16), and this was small. However, part of this may be due to the relationship between teacher-made objective tests and the amount of time spent on fractions in class (r = 0.19), which includes the lower performing classes. A closer look at the correlations between assessment tools indicated that teacher-made objective tests may have had some unique effect on achievement between classrooms. Including this variable in the HLM model resulted in a relatively moderate effect. This can be seen in Table 29.

This model accounted for 34 percent of the variance between classrooms. At this point, the additional variables were accounting for very small portions of remaining variance, less than one percent additional variance from the previous models.

Table 29
HLM Model of Student Mathematics Achievement Given Prior Achievement & Assessment Practices

Fixed Effects		*Coefficient*	*S. Error*	*T-Ratio*	*p-value*
Intercept Level-2, grand mean	γ_{00}	0.008	0.031	0.262	0.793
Prior Achievement	γ_{01}	0.233	0.022	10.474	0.000
Homework frequency	γ_{02}	0.189	0.042	4.542	0.000
Evaluation-Grading	γ_{03}	-0.184	0.060	-3.079	0.003
Teacher-made objective tests	γ_{04}	-0.081	0.038	-2.135	0.033

Random Effects		*Variance Component*	*df*	*Chi-Sq*	*p-value*
u_{0j}	τ_{00}	0.2760	319	4913	0.000
r_{ij}	σ^2	0.3641			

For final consideration at the classroom-level, the uses teachers reported for the assessment information they gathered were evaluated with respect to effects on achievement. The correlations between assessment information uses and achievement are reported in Table 30.

Table 30
Correlations between Uses of Assessment Information and Achievement Scores

Uses	*r*	Uses	*r*
Plan future lessons	-.00	Grading	-.05
Diagnose problems	-.04	Feedback	-.07

These correlations were virtually no different than zero; all correlations were below 0.10 in magnitude. It was unlikely that any of the uses of achievement information would impact residual variance in classroom achievement levels given the current HLM model. In fact, upon testing several of the uses for assessment information, none had an impact on the current HLM model.

The final model for explaining classroom achievement levels, given class-type indicators and three teacher practices, was

$$\text{Achievement}_{ij} = \beta_{0j} + r_{ij}$$

$$\beta_{0j} = 0.237\ (\text{Prior Achievement})_j + 0.19\ (\text{Homework Frequency})_j$$
$$- 0.18\ (\text{Evaluation-Grading Purpose of Homework})_j$$
$$- 0.08\ (\text{Teacher-Made Objective Tests})_j + u_{0j}\ .$$

This can be interpreted as follows: the average level of achievement in classroom j (β_{0j}) is the sum of several effects (where the average classroom math score is essentially zero), including the prior achievement or prerequisite skill of students and assessment practices. This model accounted for 35 percent of the variance in achievement between classrooms.

Relationships between Student-level Variables and Mathematics Achievement

Analyses at the student-level employing the IRT ability parameter (herein referred to as achievement score) also employed the standard error of measurement of the achievement score and weights for the student's probability of selection. This corrected estimates for two conditions, heteroscedasticity of errors inherent in IRT ability estimates and non-random sampling of students.

Student Mathematics Achievement Performance. The nature of the TIMSS mathematics assessment was described above. The scores used in the following analyses included the two-parameter logistic IRT estimated thetas and standard errors. The estimates were obtained using Multilog (Thissen, 1991a) as described above. The distribution of the resulting ability parameters was slightly positively skewed, as can be seen in Figure 31. The values were weighted for the student probability of selection and the inverse of the standard error of the ability estimates. All analyses including the ability parameter as an outcome were weighted for the student probability of selection and the standard error of the ability parameter, while all analyses on student variables were weighted for the student probability of selection as described earlier.

The precision or relative value of the assessment can be viewed in terms of the information provided at various levels of ability scores. As can be seen in Figure 32, the test provided a great deal of information in the middle of the ability scale and less at the extreme values. The figure shows two information functions, one for the two-parameter estimates (allowing the discrimination parameter to vary) and one for the one-parameter estimates (fixing the discrimination parameter for each item to some average). The two-parameter estimates provided more information at each level of the ability scale.

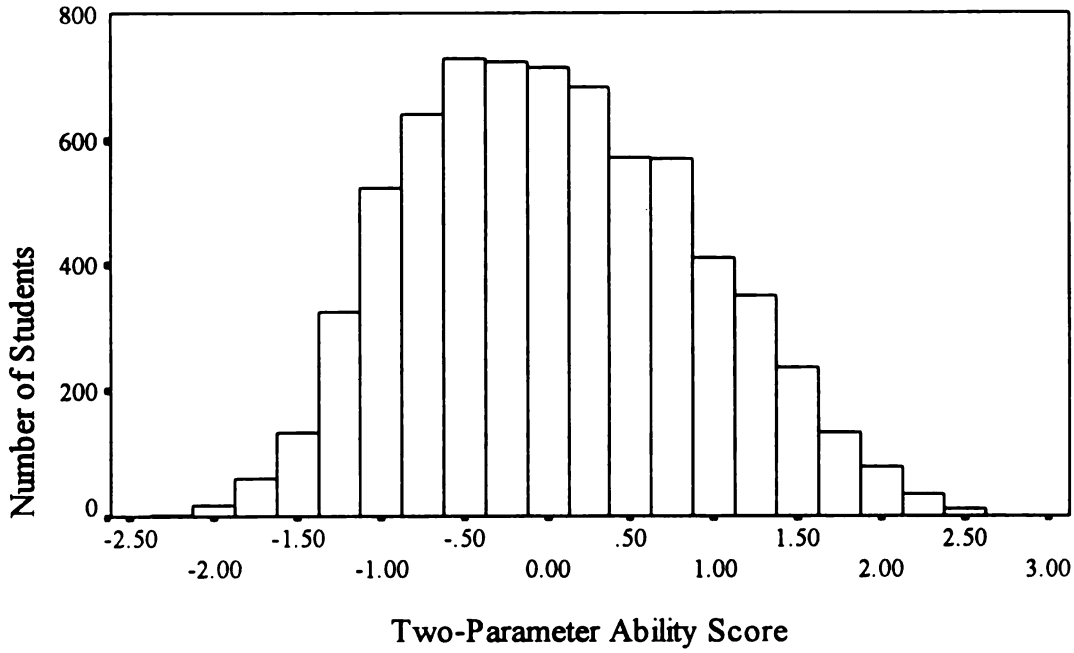

Figure 31. Two-parameter ability score distribution.

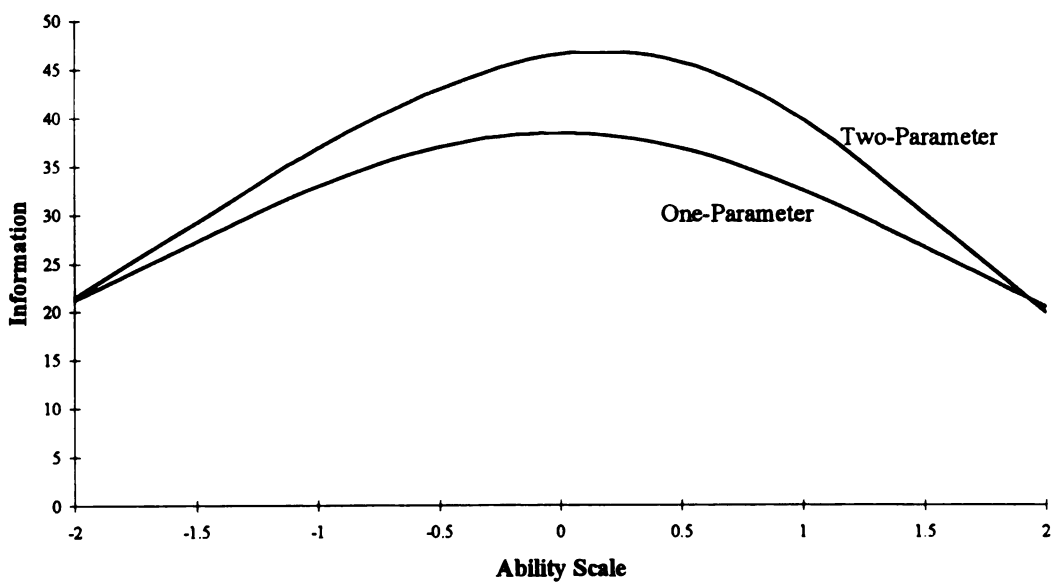

Figure 32. Information curves for the one- and two-parameter IRT models.

For analyses that have large numbers of items and where individual data were analyzed (rather than the frequency of response patterns), "there is no theoretically-justified index of overall goodness of fit of the model" (Thissen, 1991b, p.iv). However, Multilog does provide a value (-2×loglikelihood) that is approximately distributed as a chi-square if the model fits in cases involving few items (not so here). But even for large problems, the differences between these values for more or less constrained models may be treated as a chi-square (Thissen). To evaluate the potential of improved fit in using the two-parameter rather than the one-parameter model, the values of the chi-square like measure were compared. The difference between the one- and two-parameter models (-2×loglikelihood) values was 7,333 where the degrees of freedom equaled 156, the number of parameters freed in the two-parameter model (since the *a*-parameters were fixed in the one-parameter model). The resulting *p*-value for such a χ^2 (7333, 156) was less than 0.001. This suggested a significant improvement in model fit by using the two-parameter model. In addition, the marginal reliability described earlier was 0.97 for the mathematics assessment as scored by Multilog with the two-parameter model (the same marginal reliability resulted from the one-parameter model).

Although Multilog does not report fit indices for individual items, the information functions for each item were reviewed. Seven items provided little to no information at any point along the ability scale. Although these items had outlying estimated parameters, they were all essentially weighted toward zero in scoring individual abilities because they provided little to no information across the ability scale. These included seven multiple-choice items (1 fractions, 1 geometry, 1 data analysis, 2 measurement, 2 algebra) including all four types of performance expectations (knowledge, problem

solving, routing and complex procedures). These were also among the items with the lowest discrimination parameter values; all seven were below 0.4 from the range of 0.08 to 2.7 where the higher the value, the more discriminating the item.

Item difficulties ranged between -3.7 and 4.9 centered at zero with a standard deviation of 1.17 (excluding one item at 9.0). Multilog selected the location of items first, then scored individuals and placed them on the scale of the items. The weighted average ability score was 0.08 with a standard deviation of 0.92. There was no great difference between the average difficulty of items and the average ability of students.

Gender. The population was essentially half male and half female. However, in other studies of mathematics achievement, there has been a strong tendency for males to outperform females and to obtain scores with a larger variance than females in mathematics achievement tests (Bielinski & Davison, 1998). Bielinski and Davison argued that the presence of a gender difference in variability should result in a gender-by-item-difficulty interaction, for which they found empirical support in a large set of Minnesota state mathematics test scores and within TIMSS math items as well (Bielinski, personal communication, February 2, 1999).

In the TIMSS mathematics assessment results, there was no overall gender difference in mean performance; however, there was a slightly larger variance in scores for males than for females (see Figure 33). The ratio of male to female score variance was 1.08, suggesting that male scores were about eight percent more variable than were females scores. Although, the analysis of variance results suggested that the mean squares between groups was smaller than the variance within groups (no significant mean difference between genders, $F = 0.92$, $p = 0.34$), the Levene test for homogeneity of

variance, Levene's (1, 6961) = 5.69 ($p < 0.02$), did indicate a significant difference in variance between the groups. In a graphical representation similar to the one used by Bielinski and Davison (1998), the slight increase in variation among male scores can be observed: there were more males in both tails of the distribution (the solid line for males is above the dashed line for females in Figure 33 based on arbitrary score groups of equal interval length). Campbell, Reese, Sullivan, and Dossey (1996) also reported similar levels of mathematics achievement among males and females through the eighth-grade.

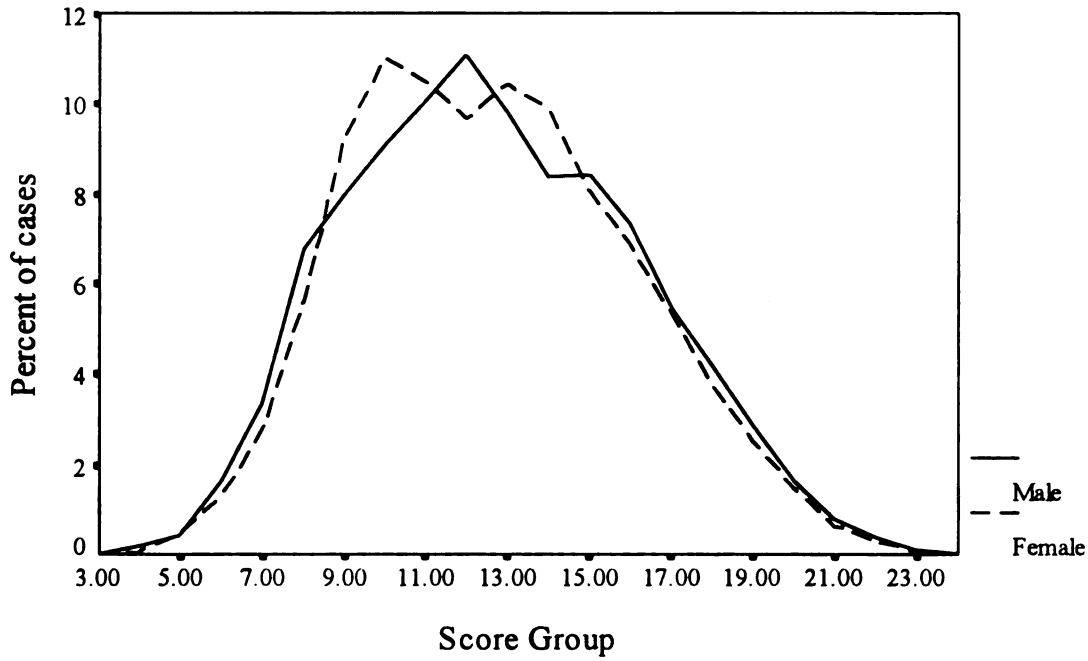

Figure 33. The percent of students in each score group of the mathematics assessment.

Grade. TIMSS targeted 13 year olds and in doing so, selected one 7th grade classroom for every two 8th grade classrooms. However, when weighted by probability of selection, the resulting sample size was fifty percent of each.

The difference in overall performance on the mathematics assessment was significant ($F = 245, p < .001$) and the variances within each grade were homogenous (Levene's (1, 6961) = 0.053, $p = 0.82$). The magnitude of the difference was 0.30, which was about 0.36 of a standard deviation on the score scale. Eighth-grade students performed just over one third of a standard deviation higher than seventh-grade students overall. This also provided an indication of the size of the score difference expected given one year of schooling at the middle school level.

English. Approximately 90 percent of the students always or almost always spoke English at home, while 10 percent did not. The difference in score performance between these two groups was significant ($F = 121, p < 0.001$) and the variances within each group were homogenous (Levene's (1, 6965) = 3.10, $p = 0.08$). The magnitude of the difference was 0.37, which was about 0.44 of a standard deviation on the score scale. Students who always or almost always speak English at home performed almost half a standard deviation higher than students who did not always speak English at home did. This was greater than the difference between seventh and eighth-grade performance, suggesting that students who did not speak English at home were performing more than a full-year lower than their English-speaking classmates.

Mother's Education. There was a wide range of experiences in terms of level of mothers' education. Table 31 displays the weighted *n*s, mean mathematics scores, and their standard deviations.

Table 31

Descriptive Statistics for Average Math Scores by Mothers' Level of Education

		Mathematics Score	
	n	*Mean*	*SD*
Finished primary school	148	-0.48	0.74
Finished some secondary school	622	-0.26	0.74
Finished secondary school	1685	-0.02	0.77
Some vocational school	540	-0.07	0.81
Some university	1681	0.12	0.83
Finished university	1571	0.35	0.93

Based on the analysis of variance results, there were significant differences in achievement scores between students whose mothers' had various levels of education; however, the variances within each group were also heterogeneous (Levene's (5, 6241) = 25.2, $p < 0.001$). In fact, the variance within each group increased as the overall performance of each group increased (from 0.74^2 to 0.93^2). The performance of students of mothers with higher levels of education varied more than students whose mothers had lower levels of education (without information regarding the quality of education). To avoid over-interpretation, the focus could be on differences between students whose mothers finished high school and those who finished college. The magnitude of this difference was 0.37, about 0.44 of a standard deviation on the score scale. This indicated that students with mothers who completed college scored more than a full year above those students whose mother only completed high school.

Mother's Expectations. Only two percent of the students reported that their mothers did not think it was important to do well in math. Twenty-six percent agreed while 71 percent strongly agreed. The differences in score performance between these groups was significant ($F = 42, p < 0.001$); however, the variances within each group

were heterogeneous (Levene's (3, 6877) = 4.27, $p = 0.005$). Since so few students reported to disagree with the idea that their mothers thought it was important to do well in math, the focus could be on the magnitude of the difference between those who agreed and those who strongly agreed. The size of the difference in average score performance was 0.13, which was about 0.15 of a standard deviation on the score scale. Students who strongly agreed that their mother had high expectations regarding math achievement scored about one-seventh of a standard deviation higher than did students who only agreed their mothers had high expectations regarding math achievement. This was about one-half the difference between seventh and eighth-grade performance, suggesting those students who reported stronger sense of their mothers' expectations scored about one-half year higher than other students.

Effort. Student effort was perhaps one of the more complex issues in these analyses. Unfortunately, as discussed above, the indicators of effort were not strong. However, students reported the amount of time they spent studying mathematics outside of school on a normal day. As previously reported, students reported spending either no time (17%), less than an hour (57%), between one to two hours (24%) or three or more hours (3%). The mean differences in math scores between groups was significant ($F = 77.4, p < 0.001$); however, the variances within each group were heterogeneous (Levene's (4, 6797) = 8.40, $p < 0.001$). Without regard to the smallest group (three or more hours), the magnitude of the average score difference between students who spent no time and less than one hour was 0.45, which was about 0.53 of a standard deviation on the score scale. Students who spent up to one hour studying on a normal day scored more than a

half of a standard deviation above those students who spent no time studying; this was equivalent to one and a half years of schooling.

In contrast, students who spent one to two hours studying math on a normal day scored about one-tenth of a standard deviation below those who spent less than one hour studying. In addition, although the group was small (Figure 34), those students who spent three or more hours studying scored more than one-half a standard deviation below students who studied less than one hour. This was likely an indicator of the need for students to do mathematics homework or to study (as well as attitude about learning math as reported above), rather than a direct linear correlate with achievement. As can be seen in Figure 35, the relationship was not linear. An improvement would include information regarding the efficiency or productivity of students in doing homework.

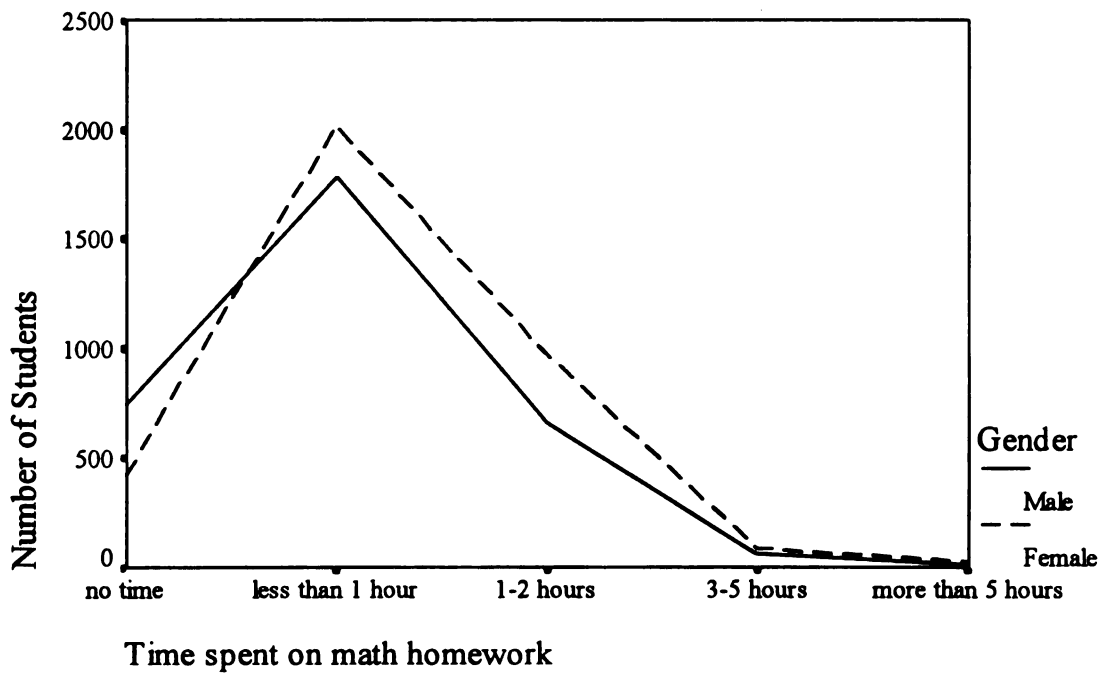

Figure 34. Distribution of time spent on homework by gender.

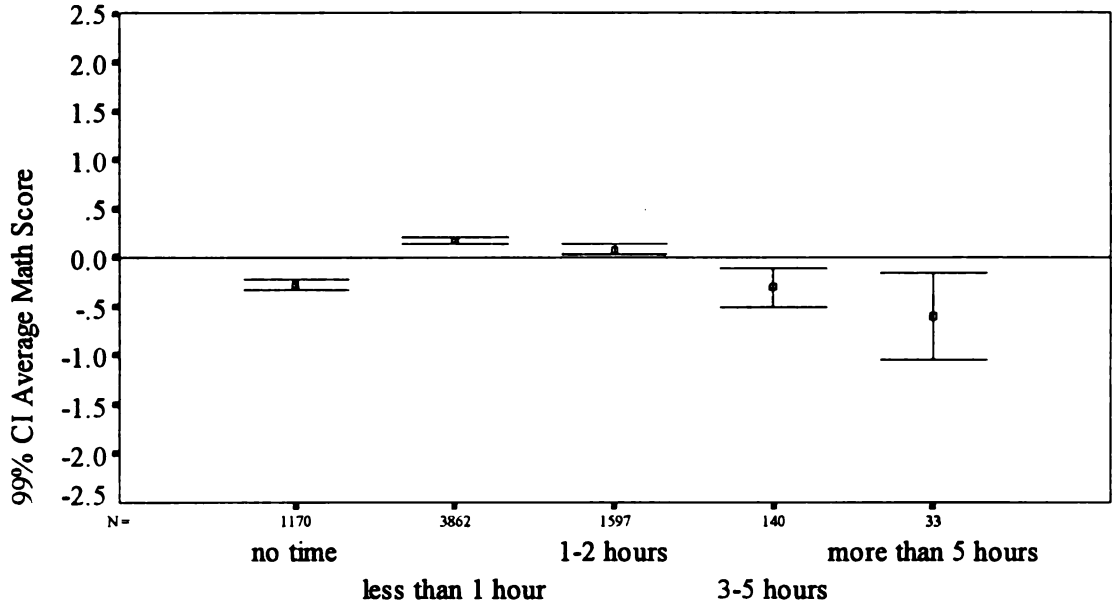

Figure 35. Diagram of mean mathematics performance by time spent doing math homework.

Self-Efficacy. Self-efficacy was described earlier and was composed of two primary elements: the students' own perceptions of potential for mastery and autonomy and their attribution of control. The attribution of control was construed simply as uncontrollable attributes (good luck, natural talent) and controllable attributes (hard work studying, memorizing notes). Thus, there were three composite indicators, including self-efficacy (potential for mastery), controllable attributions, and uncontrollable attributions.

All three of these indicators were continuous, composed of multiple indicators as described above. An evaluation of their relationship to mathematics achievement was conveniently done through an analysis of covariance. Table 32 lists the correlations among these indicators and their correlations with the achievement scores.

Table 32
Correlations of Achievement Scores and Self-Efficacy Indicators

	Achievement Score	Self-Efficacy	Controllable Attributes
Self-Efficacy (potential for mastery, autonomy)	.201		
Controllable Attributes (studying, memorizing)	-.097	.136	
Uncontrollable Attributes (luck, talent)	-.266	-.095	.077

The intercorrelations among the self-efficacy indicators were weak, as reported earlier. However, their correlations with the achievement scores were mostly as expected. Students with higher levels of self-efficacy (potential for mastery and autonomy) had higher achievement scores (r = 0.201). Students who made uncontrollable attributions regarding good mathematics performance (they attribute good performance to good luck or natural talent) had lower achievement scores (r = -0.266). Unexpectedly, the controllable attributes also had a negative correlation with achievement scores; however, it was much smaller, suggesting the possibility of no real relationship.

Combined Effects of Student Characteristics and Teacher Practices

Before adding the student-level constructs to the complete hierarchical linear model relating teacher assessment practices to classroom achievement performance at level two (described above), the student-level mediating constructs were examined using a general linear model at the student-level only. This was important to evaluate the possibility of interactions at the student-level without overburdening the HLM. All of the

student-level constructs were added to the model including all two-way and three-way interactions. None of the three-way interactions were significant, however, several two-way interactions were significant. After further evaluation, removing all non-significant two-way interaction terms, when all were simultaneously included in the full HLM model, only mother's education × uncontrollable attributions was significant. This will be interpreted below.

The full HLM model was assessed leaving all level-one slopes random except for student-level variables that were unlikely affected by teacher practices (mother's education, mother's expectations, and interaction terms). By leaving a level-one slope randomly varying at level-two, the model estimates the average slope across classrooms plus the variance of classroom slopes. By making the slope fixed, the interpretation was that the slope (coefficient) for a level-one variable did not vary across classrooms and thus its variance did not have to be estimated (it was fixed to equal zero). Each of the classroom-level variables was used to model the randomly varying slopes of level-one (classroom-level variables were used as predictors to explain variation in level-one slopes across classrooms). This is essentially a unique strength of HLM, which enables the explanation of why effects within classrooms vary across classrooms.

Three HLM models were assessed, each time removing nonsignificant terms and fixing level-one slopes to be nonrandomly varying when the corresponding variance component was nonsignificant. The third and final model was

$$
\begin{aligned}
\text{Achievement}_{ij} = {} & \beta_{0j} \\
& + \beta_{1j}\,(\text{Gender})_j \\
& + \beta_{2j}\,(\text{Mother's Education})_j \\
& + \beta_{3j}\,(\text{Mother's Expectations})_j \\
& + \beta_{4j}\,(\text{Self-Efficacy})_j \\
& + \beta_{5j}\,(\text{Uncontrollable Attributions})_j \\
& + \beta_{6j}\,(\text{Uncontrolled} \times \text{Mother's Ed})_j \\
& + \beta_{7j}\,(\text{No Homework})_j \\
& + \beta_{8j}\,(< \text{1-Hour Homework})_j + r_{ij}
\end{aligned}
$$

$$
\begin{aligned}
\beta_{0j} = {} & \gamma_{00} \\
& + \gamma_{01}\,(\text{Prior Achievement})_j \\
& + \gamma_{02}\,(\text{Homework Frequency})_j \\
& + \gamma_{03}\,(\text{Grading \& Evaluation})_j \\
& + \gamma_{04}\,(\text{T-M Objective Tests})_j \\
& + \gamma_{05}\,(\text{Average Class Self-Efficacy})_j \\
& + \gamma_{06}\,(\text{Average Class Uncontrollable Attributions})_j \\
& + \gamma_{07}\,(\text{Class \% No Homework})_j \\
& + \gamma_{08}\,(\text{Grade})_j \\
& + u_{0j}
\end{aligned}
$$

$$
\begin{aligned}
\beta_{1j} &= \gamma_{10} \\
\beta_{2j} &= \gamma_{20} \\
\beta_{3j} &= \gamma_{30} \\
\beta_{4j} &= \gamma_{40} + \gamma_{41}\,(\text{T-M Objective})_j + u_{4j} \\
\beta_{5j} &= \gamma_{50} + \gamma_{51}\,(\text{T-M Objective})_j + \gamma_{52}\,(\text{Prior Achievement})_j + u_{5j} \\
\beta_{6j} &= \gamma_{60} \\
\beta_{7j} &= \gamma_{70} \\
\beta_{8j} &= \gamma_{80} + \gamma_{81}\,(\text{Homework Frequency})_j
\end{aligned}
$$

Estimates of each of the coefficients and random effects in this model are presented in Table 33.

Table 33

HLM Model of Student Mathematics Achievement Given Classroom-Level & Student-Level Characteristics

Fixed Effects		*Coefficient*	*S. Error*	*T-Ratio*	*p-value*
Model for Classroom Means, β_{0j}					
Intercept Level-2, grand mean	γ_{00}	0.007	0.023	0.326	0.744
Prior Achievement	γ_{01}	0.119	0.019	6.336	0.000
Homework frequency	γ_{02}	0.086	0.034	2.524	0.012
Evaluation & Grading	γ_{03}	-0.121	0.052	-2.333	0.020
Teacher-Made objective tests	γ_{04}	-0.031	0.029	-1.077	0.282
Average Self-Efficacy	γ_{05}	0.262	0.091	2.875	0.005
Average Uncntrl-Attribution	γ_{06}	-0.814	0.072	-11.235	0.000
% No Homework	γ_{07}	-1.046	0.197	-5.303	0.000
Grade	γ_{08}	0.136	0.050	2.718	0.007
Models for Slopes					
Gender, β_{1j}	γ_{10}	-0.063	0.013	-4.877	0.000
Mother's Education Slope, β_{2j}	γ_{20}	0.020	0.005	4.401	0.000
Mother's Expectations Slope, β_{3j}	γ_{30}	0.028	0.012	2.335	0.020
Self-Efficacy Slope, β_{4j}	γ_{40}	0.167	0.011	14.906	0.000
T-M Objective Tests	γ_{41}	-0.029	0.013	-2.173	0.030
Uncntrl-Attributions Slope, β_{5j}	γ_{50}	-0.145	0.018	-8.047	0.000
T-M Objective Tests	γ_{51}	-0.019	0.010	-1.982	0.047
Prior Achievement	γ_{52}	0.012	0.006	2.014	0.044
Uncntrl×Mother's Ed Slope, β_{6j}	γ_{60}	0.017	0.004	4.201	0.000
No Homework Slope, β_{7j}	γ_{70}	0.077	0.021	3.700	0.000
0-1 Hr Homework Slope, β_{8j}	γ_{80}	0.120	0.015	8.032	0.000

Random Effects		*Variance Component*	*df*	*Chi-Sq*	*p-value*
Classroom Mean, u_{0j}	τ_{00}	0.1443	311	2957	0.000
Self-Efficacy Slope, u_{4j}	τ_{44}	0.0051	318	427	0.000
Uncntrl Attrbtn Slope, u_{5j}	τ_{55}	0.0028	317	408	0.001
Level-1 effect, r_{ij}	σ^2	0.3304			

This model was significantly different than the model including only classroom-level explanatory variables. It included both student-level and additional classroom-level variables. Several of the variables at the student-level were classroom-mean centered. The variables were centered around their classroom mean to retain the interpretation of the intercept: this way it remained the average classroom performance or mean performance for the classroom, where $\beta_{0j} = \mu_{Yj}$. This, however, ignored the fact that classrooms may have differed in their overall level of these variables, thus the average value for each classroom-mean centered variable was used as an additional classroom-level explanatory variable. These variables included gender, self-efficacy, uncontrollable attributions, and time spent on homework. The variables that were only modeled at the student-level included mothers' education, mothers' expectations, and the interaction of uncontrollable attributions with mother's education. These variables were viewed as unaffected by the classroom or teacher, thus they were only modeled at the student-level and grand-mean centered (rather than classroom-mean centered).

Overall, from the variance components in Table 33, the model accounted for 66 percent of the variance in classroom means. The addition of the classroom average values of several student-level variables accounted for an additional 31 percent of the variance in classroom mathematics performance. In addition, the conditional intraclass correlation (the correlation between pairs of scores in the same classroom) at this point was 0.30, reduced significantly from 0.54 in the unconditional model. This suggested that the degree of dependence among observations within classrooms that are the same on the variables included in the model was over 40 percent less. The conditional reliability of the classroom means (the reliability with which classrooms that were the same on the

conditioning variables could be discriminated from the others) was 0.87. This was less than the unconditional reliability of 0.95, which was expected since conditioning classroom performance reduced the likelihood of discriminating among classrooms reliably. Finally, the variance of the residual classroom means remained significant ($\chi^2 = 2957, df = 311, p < 0.001$); significant variation in classroom performance remained.

All of the terms in the classroom-level of the model were significant, except for the main effect of the use of teacher-made objective tests. It remained in the model because the interactions between the use of objective tests and student self-efficacy and uncontrollable attributions were significant. These will be interpreted in turn.

Student Characteristics. Gender had a significant but small slope ($\gamma_{10} = -0.06$) which suggested that females scored slightly lower than males, controlling for the other explanatory variables (i.e., all else constant). The magnitude of difference was about 0.07 of a standard deviation on the student mathematics score scale. This finding was in contrast to the bivariate analysis of gender and performance, where male scores were slightly more variable, but no mean difference existed ($t = .959, df = 6961, p = 0.338$). When student performance was conditioned on several other variables, a small gender effect did in fact result. This effect was constant across classrooms. However, gender composition of the classroom had no relationship to classroom performance.

Mother's education level and mother's expectations for performance in mathematics were both significantly positive and the effects were constant across classrooms. The effect of mothers who completed high school compared to those who completed college was about the same as the gender effect (0.07 of a standard deviation). The effect of agreeing with the statement that a student's mother expected them to do well

in math compared to disagreeing was also small but significant (0.07 of a standard deviation).

Self-efficacy had a significant positive relationship with student scores (γ_{40} = 0.167), which amounted to 0.19 standard deviations increase in math score per unit improvement in self-efficacy (on a four point scale). This effect varied significantly across ($\tau_{44} = 0.005, p < 0.001$). The self-efficacy effect was dependent on the level with which teachers used teacher-made objective tests (γ_{41} = -0.029). The self-efficacy slope (0.167) was reduced by 0.029 for each level of use of objective tests from none (0) to a great deal (4). Another way to interpret this cross-level interaction is that self-efficacy had a stronger relationship with math scores in classroom where teachers did not heavily use teacher-made objective tests. For reasons to be explained in the discussion below, the use of teacher-made objective tests had a negative effect on the positive impact of a strong sense of self-efficacy.

The use of uncontrollable attributions by students (attributing success in mathematics to luck and natural talent) had a negative relationship with math scores, all else constant (γ_{50} = -0.145), which varied across classrooms ($\tau_{55} = 0.0027, p = 0.001$). This amounted to a 0.17 standard deviation reduction in math scores for each unit of uncontrollable attributions made by students (on a 5-point scale). This was also effected by the use of teacher-made objective tests (γ_{51} = -0.019) and relative prior achievement level of the class (γ_{52} = 0.012). The use of teacher-made objective tests strengthened the negative relationship between uncontrollable attributions made by students and performance, while the prior achievement level of the class weakened the negative

relationship. For students in classes with higher prior math achievement, the use of uncontrollable attributions had less of a negative relationship with math scores.

Since self-efficacy and uncontrollable attributions varied randomly across classrooms (τ_{44} and τ_{55} were both significant), their intercorrelations were evaluated. Table 34 contains the intercorrelations for all three random effects, including the intercept. The correlation between effects for self-efficacy (β_{4j}) and uncontrollable attributions (β_{5j}) was moderate (τ_{54} = -0.542), indicating that classrooms where self-efficacy had a larger effect were also classrooms where uncontrollable attributions had a smaller effect, likely due to some common causes. One common cause, as reported here, was teacher's use of objective test (and possibly other classroom-related practices). Correlations with the intercept were small, which indicated that the effects of self-efficacy and uncontrollable attributions were not related to mean classroom performance.

Table 34
Intercorrelations of Random Effects (Tau)

	Intercept, τ_{00}	Self-Efficacy, τ_{44}
Self-Efficacy, τ_{44}	0.151	
Uncontrollable Attributions, τ_{55}	0.207	-0.542

There was a significant but very small interaction between the use of uncontrollable attributions and mother's education level (γ_{60} = 0.017). This suggested that effect of uncontrollable attributions on math scores depended on mother's education level. This relationship is displayed in Figure 36. The slope between uncontrollable

attributions and math scores is steeper for students whose mothers completed college and flatter for mothers with little education. Also, it appears that education level has a larger impact on average math scores of students who rarely use uncontrollable attributions but little impact on average math scores of students who often use uncontrollable attributions. These things seem to be at work simultaneously.

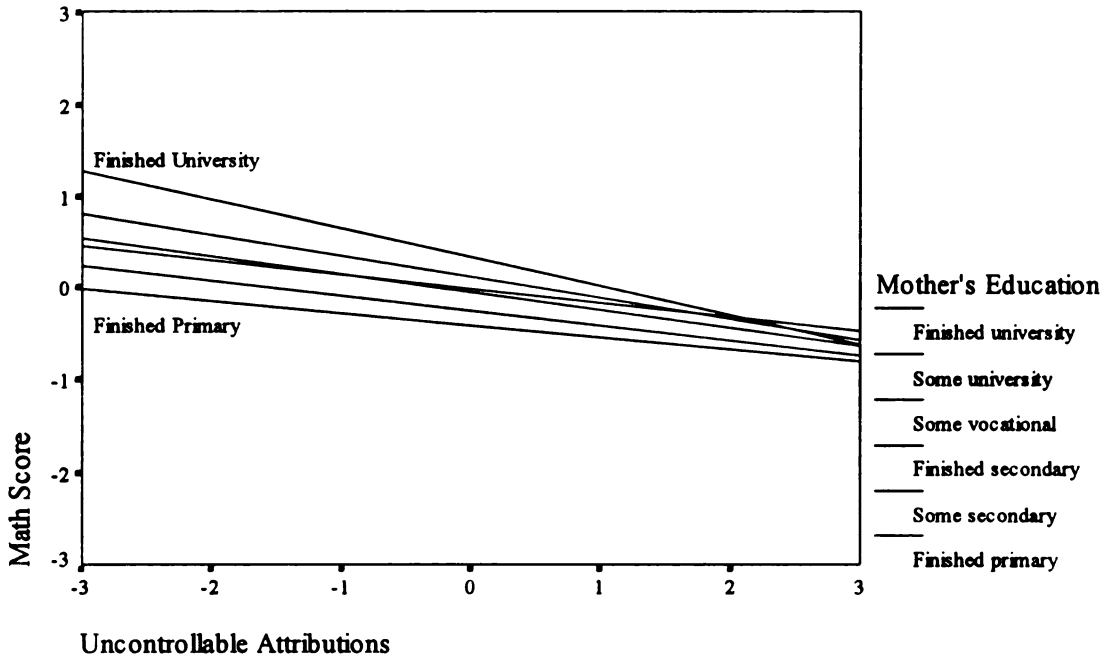

Figure 36. Interaction effect between the use of uncontrollable attributions and mother's education level.

Finally, the amount of time students spent on homework had a significant relationship with math performance, all else constant. Students who did no homework each day performed slightly higher on average than those students who spent more than one hour a day on math homework ($\gamma_{70} = 0.077$). Again, as described above, the students

who spent more than one-hour each day studying mathematics were likely students who essentially needed to study more because of poor performance. Students who did about one hour of homework each day scored at an even higher level on average ($\gamma_{80} = 0.120$), about 0.14 standard deviations above the mean. These effects did not vary across classrooms and was unaffected by the level of homework assigned as reported by the teacher.

Classroom Characteristics & Teacher Assessment Practices. The relative prior math achievement indicator was a significant explanatory variable for classroom-level performance ($\gamma_{01} = 0.119$), all else constant. Students enrolled in classes with highest relative prior achievement scored about 1.46 standard deviations above students in classes with the lowest relative prior achievement (where prior achievement ranged from -4.0 for remedial classes to +4.0 for algebra classes). In addition, grade had a significant impact ($\gamma_{08} = 0.136$) as expected. Students in eighth-grade classrooms scored about 0.21 standard deviations above the seventh-grade average classroom performance, all else constant.

The frequency with which teachers assigned homework (simultaneously including teachers who more frequently assigned text-book problem sets) had a significant relationship with math scores ($\gamma_{02} = 0.086$), all else constant. The difference in classroom performance for those classrooms where teachers assigned homework every day, compared to teachers who assigned homework once a week, was 0.40 standard deviations higher in terms of average classroom performance.

Frequent use of assessment information for the purpose of evaluating and grading students, without a primary use for direct feedback, had a negative relationship with

classroom performance (γ_{03} = -0.121), all else constant. The difference in classroom performance for those classrooms where the teacher sometimes used assessment information for grading purposes versus always did so was 0.19 standard deviations. Similarly, the use of teacher-made objective tests had a small but negative relationship with classroom performance (γ_{04} = -0.031), all else constant. The difference in classroom performance for those classrooms where the teacher never used teacher-made objective tests versus those that did so quite a lot was 0.14 standard deviations.

Three student-level characteristics that had significant relationships to math scores within classrooms also significantly explained variation in classroom mean performance, all else constant. The average self-efficacy level of the classroom (γ_{05} = 0.262), the average level of uncontrollable attributions made by students in a classroom (γ_{06} = -0.814), and the percent of students in the classroom who usually did no homework (γ_{07} = -1.046) were significant explanatory variables at the classroom-level. After accounting for these differences among students within classrooms, their average effect on classroom performance remained significant.

Briefly, the largest impact these variables had could be presented in terms of a class with the lowest average value and a class with the highest average value on each variable, all else constant. This comparison would lead to a maximum effect size of 0.68 standard deviations improvement in average class math performance due to overall classroom positive self-efficacy, 2.65 standard deviations improvement in average class math performance due to fewer overall classroom uncontrollable attributions, and 1.25 standard deviations drop in average class math performance due to doing ***no*** homework.

Assessing the Adequacy of the Hierarchical Linear Model

The validity of inferences based on linear models relies on the defensibility of the assumptions of the model. In HLM, these assumptions include specification assumptions at both levels required by ordinary least-squares procedures for the structural part of the model and assumptions regarding the distribution of errors at both levels for the random part. There are five key assumptions for a two-level HLM (Bryk & Raudenbush, 1992).

Distribution of Error at Level-One. Each r_{ij} (student *i* deviation from classroom *j* mean) is independent and normally distributed with a mean of zero and constant variance σ^2 across students within classrooms, in effect, $r_{ij} \sim N(0, \sigma^2)$. Figure 37 contains a normal probability plot of the level-1 residuals and provides evidence of a fairly normal distribution, excluding slight deviation in the tails.

The HLM program provided a test for the variance homogeneity assumption. The within-classroom variances were heterogeneous across classrooms ($\chi^2 = 386$, $df = 311$, $p = 0.003$), violating the constant variance assumption. However, the expected impact on the estimation of parameters and standard errors was minimal. Kasim and Raudenbush (1998) recently reported that the restricted maximum likelihood pooled estimate of the variance (as computed by HLM) compensated for heterogeneity by increasing in size. These estimates remained unbiased, although may not have been asymptotically efficient, and for large numbers of level-2 groups, "standard errors will not be sensitive to heterogeneity of variance" (p. 108). In their study, the large numbers of groups condition included 100 observations; in this data set, there were 328 classrooms at level-2.

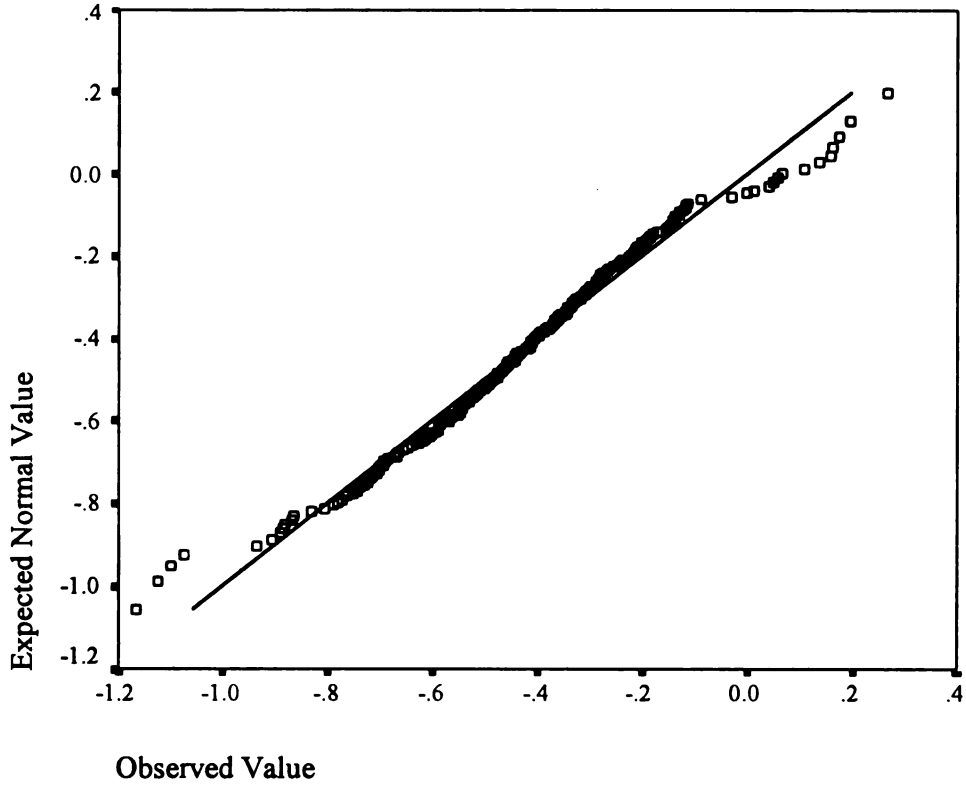

Figure 37. Normal Q-Q plot of level-one residuals (mdrsvar).

Independence of Explanatory Variables and Error at Level One. The explanatory variables are assumed to be independent of the error terms at level one, in effect, $Cov(X_{qij}, r_{ij}) = 0$ for all q (explanatory variables). When an explanatory variable is correlated with the error term, there are likely confounding variables since their influence is included in the error term. These excluded variables are assumed independent of the explanatory variables in the model. This assumption was assessed through a thorough examination of bivariate relationships among numerous potential exploratory variables not included in the final model. The null correlation between predicted values and residuals at level one provided some evidence of appropriate specification at level one.

The existence of possible confounding variables not included in this data was not testable; however, this should be one of the purposes of additional research.

Distribution of Error at Level Two. The vector of random errors at level two are multivariate normal, each with a mean of zero, some variance τ_{qq} and covariance $\tau_{qq'}$. This model included three residual classroom effects, classroom means (u_{0j}), self-efficacy slopes (u_{4j}), and uncontrollable attribution slopes (u_{5j}). A Q-Q plot of the Mahalanobis distances assessed the multivariate normality of these random errors, as displayed in Figure 38. Bryk and Raudenbush (1992) suggested that if the normality assumption is true, then the Mahalanobis distances should be distributed approximately $\chi^2(3)$. With the exception of one outlier, the multivariate normal distribution assumption was tenable.

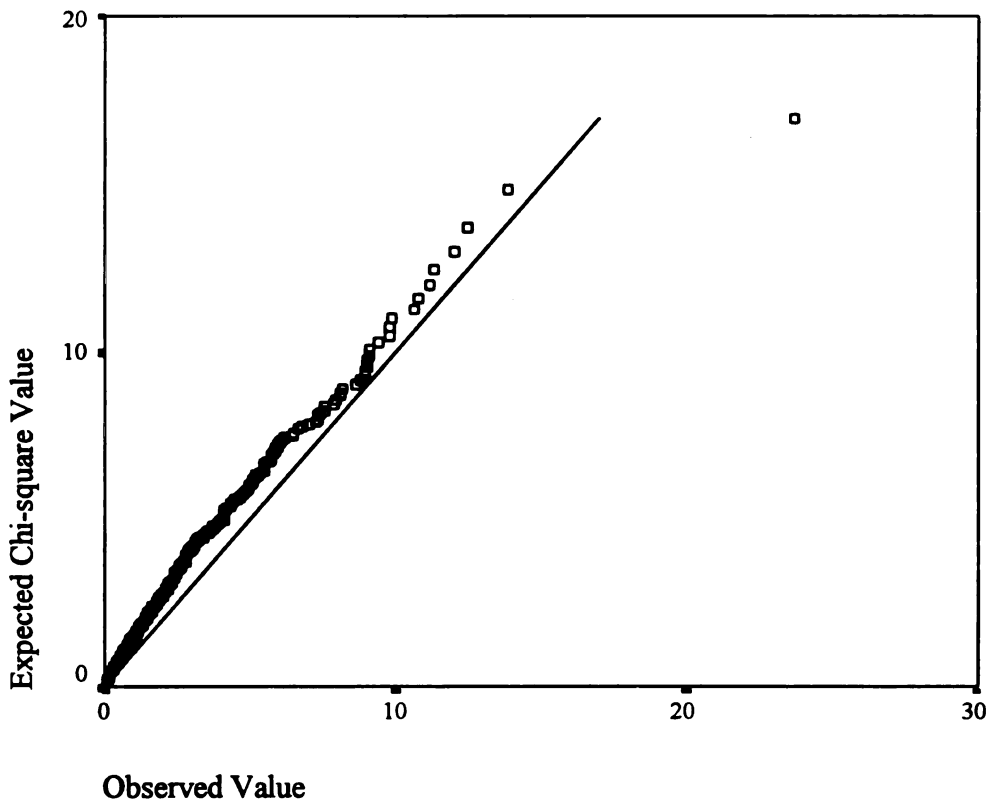

Figure 38. Normal Q-Q plot of level-two residuals (Mahalanobis Distances).

Independence of Explanatory Variables and Error at Level Two. The explanatory variables are assumed to be independent of the error terms at level two, in effect, $Cov(W_{sj}, u_{qj}) = 0$ for all *s* (explanatory variables). Again, this is an assumption regarding specification and lack of confounding variables. Several potential level two explanatory variables were evaluated in terms of their relationship with significant explanatory variables included in the model and the outcomes. This assumption is properly evaluated based on theory and for those classroom characteristics that were not included in this data, future research will play an important role in further specification of the mode. One additional indicator were the null relationships (i.e., nonsignificant correlations) between the residuals for the three random components (i.e., classroom means, self-efficacy slopes, and uncontrollable attribution slopes) and their fitted values (predicted values). The lack of relationship between residuals and predicted values provided some evidence of appropriate specification.

Independence of Errors between Level One and Level Two. The errors at level one and level two are assumed to be independent, in effect, $Cov(r_{ij}, u_{qj}) = 0$ for all *q* random components. A scatterplot of the residuals from level one and the Mahalanobis distances for the multivariate estimate of residuals at level two provided evidence to support this assumption, seen in Figure 39. With the exception of one outlier, as in Figure 42, there was no discernible relationship between the residuals from both levels.

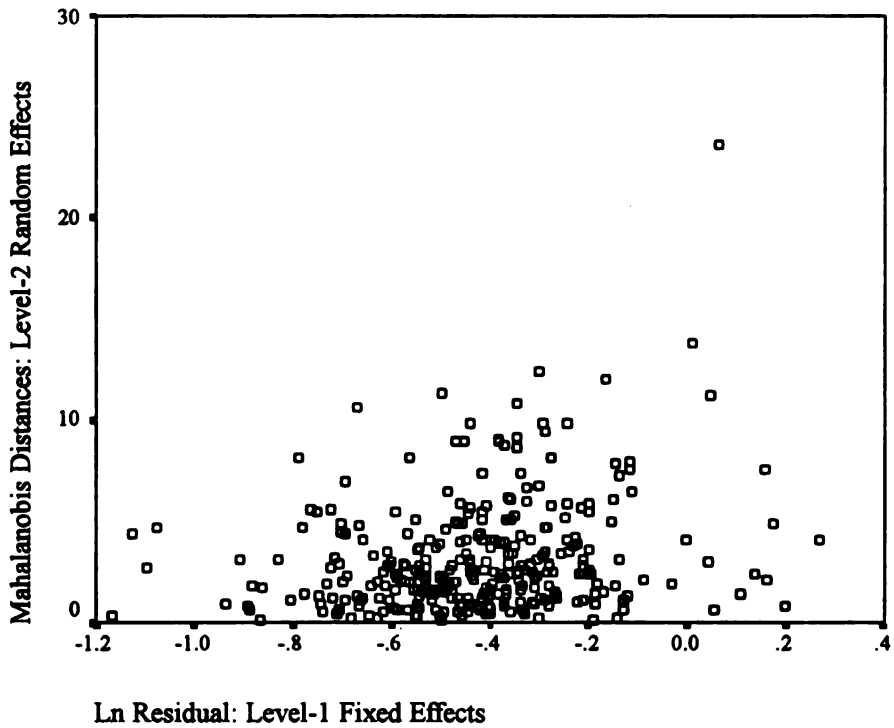

Figure 39. Scatterplot of residual measures from level one and level two.

Based on the evaluation of the five key assumptions in formulating a two-level HLM, the evidence supported the adequacy of the model and, essentially, the appropriateness of the inferences regarding the parameter estimates and the fit of the model to the data.

Classification of Teacher Assessment Practices

Above, hierarchical cluster analysis was used to classify teacher assessment practices. This was done primarily to summarize correlated practices for efficient inclusion in a larger linear model relating practices to mathematics performance. However, other more flexible options were available to conduct additional classification procedures, including discriminant analysis.

Discriminating Ability of Assessment Practices. Briefly, discriminant analysis was used to evaluate the degree to which teacher characteristics could be used to identify a teacher's use of particular assessment tools in a reliable manner. Several teacher characteristics were used to assess the ability of the assessment practice variables to identify teachers based on their known characteristics, including level of algebra taught, frequency of homework assigned, and gender. The results from the discriminant analyses were unsatisfactory. Most of the classification results were poor, where less than 50 percent of the original grouped teachers were correctly classified. One interpretation of these results was that, overall, teachers' assessment practices on the whole differed little by type-of-class indicators, frequency of homework assigned, and gender.

CHAPTER V

Summary & Discussion

This dissertation began with an account of recent (at the time) media attention on assessment and accountability, particularly in Michigan. One year later, the attention has not waned, and in some respects, has been magnified not only in Michigan, but also nationally (Sack, 1999). In his 1999 state of the union address, President Clinton stated that he would "send to Congress a plan that, for the first time, holds states and school districts accountable for progress and rewards them for results" (Sack, p. 21). Among his proposals was to hold states and school districts responsible for the quality of their teachers.

The primary research question used to direct this work is reviewed here explicitly: What are the interrelationships between classroom practices, student characteristics, and achievement performance? There is also discussion regarding policy implications and recommendations for teacher training and professional development. The chapter concludes with a framework for a comprehensive research program in these areas with recommendations for future investigations.

Overall, the assessment practices of teachers were complex and not easily characterized. The use of homework tasks and various other assessment tools, as well as the purposes for which these tools were used, were multifaceted. In short, classroom assessment practices were directly related to student performance and interacted in unique ways with student characteristics.

Relationships between Classroom Practices, Student Characteristics, and Performance

Again, the level of inference that is appropriate for these data is at the student-within-classroom level. The data are based on a representative sample of students and their mathematics teachers. Given the unconditional HLM model (without any explanatory variables), 54 percent of the variance in student performance scores was between-classrooms while 46 percent was within-classrooms. This was a significant finding in that more than half of the variance in mathematics performance could be attributed to classrooms. Unlike most HLM studies where schools are the level-two grouping structure, mathematics classrooms are likely more homogenous in terms of student performance (i.e., skill level) than are schools.

The full HLM model with student and classroom level explanatory variables explained 65 percent of the variance between classrooms and an additional 8 percent of the variance within classrooms. Recall that the within classroom variance was based on the original partitioning of variance into within and between classrooms. Once the variance in student performance was partitioned, 46 percent remained within classrooms. Of this, the HLM model explained an additional 8 percent of the within-classroom variance. The primary objective of this project was to identify characteristics at the classroom level that explained significant variance in classroom performance. The focus of the remaining discussion will rely on the classroom level explanatory characteristics.

The explanation of 65 percent of the between classroom variance was a significant result. The largest amount of variance was due to the type of math class indicators (fractions and algebra). As expected, identifying the type of math class students were in was a significant contributor to explaining variation in classroom

performance. Students were likely grouped into classes appropriately based on prerequisite skill level, which likely correlated with ability. However, some classroom performance variance remained. The additional classroom characteristics contributed significantly to this remaining variance.

Several student-level characteristics also differed on average across classrooms and contributed to the explanation of between classroom variance in performance. The most significant characteristic was the average level of uncontrollable attributions made by students in a classroom. This had a significant negative relationship with classroom performance, as expected. On the other hand, the average level of self-efficacy of the classroom had a significant positive relationship with classroom performance, although not as much of an impact as with the uncontrollable attributions. These are areas where teachers have a potential to affect students in terms of developing self-efficacy regarding potential for mastering mathematics and the level of uncontrollable attributions students make in the classrooms (Brookhart, 1997; Glasser, 1985; Marsh & Craven, 1997; Wigfield, Eccles, & Rodriguez, 1998).

Classrooms where large proportions of students did no homework were also classrooms where teachers assigned less homework. However, even though classrooms where teachers assigned frequent homework also had smaller proportions of students that did no homework, they also had students who did homework anywhere from none to more than three hours a day. Overall, both of these characteristics had a significant independent effect; more frequent homework was associated with higher performing classrooms and larger proportions of students who did no homework were associated with lower performing classrooms.

Frequent homework assigned by teachers also improved the effect of doing about one hour of homework each day. That is, in classrooms where teachers assigned more homework, the positive effect of students doing homework about one hour on an average day was greater, all else equal. This indicated that students who studied one hour each day performed at a higher level when teachers actually assigned more homework--a result that may be confounded with efficiency of homework completion. Again, this is an area that teachers may have some control over. The range in percent of students in a classroom that did no homework was from zero to 78 percent. The resulting difference in classroom performance, all else constant, was more than one standard deviation on the classroom average score scale.

These findings are in partial agreement with previous research as reviewed above. Although there is some evidence that eighth-grade students spend time on homework each day (Walberg, 1991), the amount of time spent was not always clearly related to achievement. Cooper (1989), Keith et al. (1993), and Keith and Cool (1992) did find significant relationships between homework and achievement, however, only Cooper, in his review of research, suggested that this may be curvilinear.

Finally, certain classroom assessment practices were significantly related to classroom performance, controlling for all of the above characteristics. Although the frequency of homework had a positive relationship to performance, this was also highly related to the use of textbook problem sets as homework activities. Reliance on textbook problem sets could also indicate a reliance on textbook-based instruction, which may ultimately relate to strong performance on objective assessments such as TIMSS. Neither

frequency of homework nor reliance on textbook problem sets were related to the level of algebra content in the class, so it appears independent of the type of math class.

The use of homework for grading and evaluation of students had a negative relationship with classroom performance, all else constant. This was possibly due to the low level of interaction with students based on their homework performance in classrooms where the primary use of assignment results was for grading. Although feedback to students had a positive bivariate relationship with classroom performance ($r = 0.12$), this effect was not significant in the presence of the other variables (algebra-focused classrooms had teachers who were slightly more likely to use homework as an opportunity to provide feedback to students).

The use of teacher-made objective tests also had a negative relationship with classroom performance. As mentioned earlier, the difficulty teachers face in developing high quality objective tests may have influenced this result. Unfortunately, measures of the quality of teacher constructed objective tests were not available. The use of low-quality teacher-made objective tests could result in lower performance on large-scale objective tests in a number of ways. Low quality tests do not provide reliable indicators to students regarding their achievement.

It is difficult to assess whether these results concur with previous research because of the varying definitions of assessment tools and uses throughout the literature, and because of the absence of research investigating the relationships between assessment practices and student performance. Stiggins, in two separate studies (Stiggins & Bridgeford, 1985; Stiggins & Conklin, 1992), found higher use of teacher-made objective tests than were reported here. Clearly, in both cases, teacher-made tests were

more common than published or other written tests. Salmon-Cox (1980) also found that teachers rely on their own tests more than on interactions with students or homework. None of these results were evaluated vis-à-vis performance.

Similarly, results on the uses of assessments are difficult to compare to previous research because of definitional differences. Stiggins and Conklin (1992) and Stiggins and Bridgeford (1985) reported that teachers use assessments largely to assign grades. Although they found that eighth-grade teachers use teacher-made objective tests for diagnosis, grouping, evaluating, and reporting; math teachers rely more on teacher-made objective-tests rather than performance assessments for grading, not for diagnosis.

As stated earlier, teachers communicate learning objectives through their assessments as well as indicate to students content and skills that they believe are important. If these things are poorly communicated, a likely result is poor performance. Poorly designed objective tests can also result in confusion among students in terms of their understanding test questions and ultimately their understanding of important concepts. As some have argued, students may learn as much from taking tests as any other activity they engage in; although most expect learning to primarily occur prior to testing. This assumes congruence between the tests as constructed by teachers and the instructional learning goals, which is often not achieved among middle school mathematics teachers (McMorris & Boothroyd, 1993). When evaluating instructional effectiveness, the fit between classroom assessment instruments and curriculum must also be evaluated. Similarly, when evaluating classroom assessment instruments, how well they encompass instructional learning goals is an important consideration, particularly if assessment instruments are to contribute to those instructional learning goals as well.

The importance of each test item in terms of the content on which it is based and the cognitive behavior required to correctly answer it requires the item to be written without flaws (Haladyna, 1994, 1997). In addition, the limited training of teachers in educational measurement in general and item-writing specifically has been well-documented (Plake & Impara, 1997).

The impact of low-quality teacher-made objective tests is still an area that requires careful attention. This is particularly important in terms of the call for classroom assessment reform and the predominant use of objective formats for large-scale testing programs adopted by most states--whether they be low or high stakes. However, at this point, the quality of the teacher-made objective tests by the TIMSS teachers is unknown, but likely similar to other findings reported above.

To complicate matters even more, high reliance on teacher-made objective tests as an assessment tool in middle school mathematics classrooms had a negative relationship with the effect of self-efficacy (i.e., the self-efficacy slope across classrooms) and a positive relationship with the effect of uncontrollable attributions at the student-level (i.e., the uncontrollable attribution slope), all else constant. For students in classrooms where teacher-made objective tests were prevalent, the positive effect of self-efficacy was weaker than in classrooms where teacher-made objective tests were not prevalent. Loosely speaking, greater focus on teacher-made objective tests neutralized the positive effect of self-efficacy and strengthened the negative impact of uncontrollable attributions on performance. There was evidence to suggest that the use of teacher-made tests as an assessment tool had indirect as well as direct negative relationships to student

mathematics performance. A third, possibly confounding factor, as discussed earlier, could be quality of teacher-made objective tests.

These are areas where some research is being done, but careful attention to outcomes could inform this work a great deal. The unique findings of interactions that crossed levels also deserve additional attention (i.e., use of teacher-made tests at the teacher-level moderated the effect of self-efficacy and uncontrollable attributions at the student-level; homework frequency as assigned by teachers moderated the effect of students doing about one hour of homework a day). This suggested that unique combinations of teacher practices and student characteristics yield different results in terms of middle school mathematics performance.

As argued earlier, uncovering these unique combinations may help lead to the most informed policy making regarding assessment reform efforts and determining appropriate teacher training and professional development activities. The issues related to the use of assessments in the middle school mathematics classroom are certainly complex. The analyses here only provide some indication of that complexity. With improved measures of these important student, teacher, and classroom level characteristics, a clearer portrait of the complex demands of classrooms and their assessment environment can be developed. The results of these correlational analyses do not provide evidence to support causal inferences. However, lack of causal evidence has rarely prevented the design of educational policy. At this point, evidence regarding the complexity of the classroom assessment environment does not extend far beyond what was presented earlier. This study added considerably at least to the level of complexity in considering the role of assessment practice, if not to defining some of the relationships

involved in assessment practice. Any future research in these areas should attempt to capture the complexity inherent in assessment systems and their effects across various levels of the education system.

Informing Public Policy

As John Brandl (Dean of the Humphrey Institute of Public Affairs at the University of Minnesota and two-term member of the Minnesota State Senate) has frequently stated, what we know rarely informs what we do. The task of the educational researcher and policy analyst is to craft research findings in a way that is amenable to policy development and to craft policy statements that are amenable to implementation. Most policy decision-makers look for hard evidence to guide their decisions. Relevant decisions in this arena include teacher and administrator licensure and certification requirements, and school and district assessment programs and related policies. This implies subsequent consideration of the requirements for teacher education programs and subsequent professional development activities. Evidence to inform these decisions must be able to illuminate the consequences of specific practices, based on the evaluation of some prescriptive theory. Much of the educational research underway today consists of the search for best practices. However, the consequences of assessment practices are exceedingly difficult to isolate, since learning occurs both within and outside the classroom and classrooms are diverse in terms of content and student and teacher characteristics and experiences.

In a review of the evidence regarding the consequences of assessment (primarily large-scale assessment), Mehrens (1998) did not uncover evidence suggesting that large-

scale assessment programs have impacted classroom assessment practices. There may be some evidence to suggest that high stakes assessments may impact curriculum and instruction, but Mehrens also suggested that professional development is likely to have greater impact in these areas. So it appears that professional development designers and teacher educators, as well as teachers and administrators, are the likely recipients and primary audiences for the results of investigations regarding classroom assessment practices and their impacts. This is consistent with much of the literature and recommendations of many of the researchers whose work was presented above. Much of the work reported above was done in an applied context with the hope that their work will have practical implications for educators and policy makers.

Before the presentation of the future research framework, a formal consideration of the nature of educational research is important.

The Nature of Knowledge in Educational Research

Much of the results presented above were descriptive in nature. As mentioned earlier, no attempt was made to present findings to support causal arguments. In fact, the nature of most educational research is such that the accumulation of consistent claims, causal or otherwise, is difficult to achieve. Educational research has been characterized as producing soft knowledge rather than hard knowledge and producing applied knowledge rather than pure knowledge (Labaree, 1998).

> Two characteristics in particular make it difficult for researchers in soft knowledge fields to establish durable and cumulative causal claims. One is that, unlike workers in hard knowledge fields, they must generally deal with some aspect of human behavior. This means that cause only becomes effect through the medium of willful human action, which

> introduces a large and unruly error term into any predictive equation. (Labaree, 1998, p. 5)

Even when researchers try to measure and account for those elements of human nature they believe to affect willful action (e.g., self-efficacy into effort), unique personal characteristics that may manifest themselves in teaching and learning, curriculum, school governance, the organization of schools, and reform efforts all disable attempts by educational researchers to develop a line of causal claims that are "verifiable, definitive, and cumulative," as Labaree (1998) submitted. He also argued that

> the only causal claims educational research can make are constricted by a mass of qualifying clauses, which show that these claims are only valid within the artificial restrictions of a particular experimental setting or the complex peculiarities of a particular natural context. Why? Because the impact of curriculum on teaching or teaching on learning is radically indirect because it relies on the cooperation of teachers and students whose individual goals, urges, and capacities play a large and indeterminate role in shaping the outcome. (p. 5)

This leaves the majority of work of educational researchers to be applied, often to very limited and particular settings. Generalizability is a serious issue. Labaree also implied a use for the results of educational research; one based on the normative nature of applied research to improve outcomes. However, this recognition does not preclude the continuing work of thousands of educational researchers. No one has suggested that it is possible to uncover the determining elements of individual achievement or achievement at the classroom level.

As suggested previously, the unique combinations of a complex system are what drive outcomes. This was one of the attractive elements of the TIMSS. The classrooms were diverse, from all regions of the country, composed of students with various

backgrounds, and an accompanying rich background questionnaire and comprehensive mathematics assessment. Of course not all of the student, teacher, or classroom characteristics of interest were measured equally well. The promise of such a rich database is in its ability to provide indicators of the levels of complexity, levels which can be further refined and uncovered through additional research.

With these issues in mind, the following suggestions for future research are offered to provide a framework for a more comprehensive research agenda. The research framework here was developed with application in mind, particularly regarding the development of appropriate reform policies, but with the interest in helping teachers and students as well.

A Framework for Research on Classroom Assessment Practices

Much of the work needed in this area could be considered theory construction or more generally, model building. Just as scientific knowledge cannot explain why things exist the way they do, researchers can at least provide descriptions of how events are related. A well-constructed theory can provide (1) a classification scheme, (2) sensible explanations or predictions, (3) an awareness of comprehension or understanding of certain phenomena, and (4) the possibility or capacity for control of certain phenomena (Reynolds, 1971). These are somewhat high expectations at this point. However, it is not unreasonable to expect quick resolution of the first two provisions, including a classification scheme for classroom assessment practices and sensible explanations of the events commonly experienced in the classroom assessment environment.

This framework for future research briefly outlines considerations for model building, construct identification and measurement, estimation of relationships, and inference with an eye toward the validity or relevance of interpretations. This closely follows a common statistical paradigm, consisting of model specification, parameter estimation, and assessment of model-data fit. Unfortunately, the third step in testing fit is often seen as the least important or ignored completely. It is, in fact, potentially more important than the estimation of parameters (William Schmidt, personal communication, January, 1996).

Model-Building. In general, the model presented by Brookhart (1997) remains an important and appropriate model. It was based on an appropriate review of the literature and various components have been supported empirically in the past. Based on the results of this project, several components were also supported (i.e., significant relationships between assessment activities, self-efficacy, student effort, and achievement performance). One initial step in the continuance of this line of research should include additional evaluation of this model. This is particularly important since students' perception of the significance of feedback was not included in this study and student effort was not effectively measured. In addition, the model should be evaluated in terms of its current recursive state (i.e., all of the paths are in one direction). It is possible that positive performance, relatively speaking, motivates effort as much as positive effort translates into improved performance.

The addition of a direct relationship between classroom assessment practices and performance should be included and continue to be evaluated. These must be evaluated

through careful specification of hierarchical linear models, which are appropriate for testing interactions between levels of nested data.

One possibility for improving the model building stage is to investigate the possibility of using meta-analysis as a way to link studies based on parts of the larger model. The use of meta-analysis for theory building is beginning to gain recognition, but has not been widely adopted (Becker & Schram, 1994; Eagly & Wood, 1994; Miller & Pollock, 1994). If sufficient studies exist which estimate several coefficients (relationships) of the larger model with sufficient overlap, it is possible to piece together complete relationships through multivariate synthesis of the pieces. It may not be necessary to obtain indicators on all constructs in a single comprehensive study to complete the evaluation of the full model. Based on several reviews, it may be possible to conduct such a synthesis (Brookhart, 1997; Cooper, 1989; Crooks, 1988; Dempster, 1997; Keith & Cool, 1992).

Construct Measurement. This may be the primary focus for some time, given the complex nature of the constructs themselves. The following constructs are important and require theory-driven operationalization and careful measurement. For several of these constructs, reliable measures have not been developed and perhaps may never be (e.g., invested mental effort by students).

Classroom assessment practices must be clearly operationalized, identified, and measured. In this project, assessment practices have been characterized by two facets and two dimensions. Two facets include (1) various homework activities and (2) other assessment tools. Each is composed of two dimensions, including (1) the types used and (2) how the resulting information is used, or what teachers do with the information once

obtained. The tasks included here were in part selected because of the availability of information in the TIMSS. Additional tasks could be added as well. Important in the measurement of tools and uses will be indicators of quality and frequency. It will also be important to know if uses of assessment information differ for different assessment tools.

An effort to provide a standard for describing assessment practices must be made. One problem in comparing results across studies, regarding the prevalence of certain practices, lies in the diversity of labels and definitions used by researchers of assessment practices. For example, some researchers delineate between teacher-made multiple-choice and constructed-response tests while others have reported on teacher-made tests as a whole. This hinders attempts to accumulate information, track trends, or make comparisons. In addition, as researchers design instruments to gage how teachers use assessment tools, these must be in relation to specific tools; knowing how teachers use assessment information in general is informative, but the real questions have to do with how different tools or assessment practices are used for different purposes.

Further work could be done to develop a classification scheme or identify profiles of assessment practices. Multidimensional scaling could provide a strong tool to classify teacher classroom assessment practices to identify major profiles of those practices. Correspondence analysis between individual profiles and the major profiles could illuminate important aspects of assessment practices that may ultimately hold stronger promise for uncovering relationships between assessment practices and student performance.

One aspect missing from the current model is related to assessment quality. This is potentially a very important determinant of the effectiveness of assessment tools as

used by teachers. This is an area where experimental or quasi-experimental settings would provide convincing results. Until quality indicators become available, the nature of the relationships between reliance on certain assessment tools and student self-efficacy and performance in the subject matter will continue to be part supposition, again relying on pieces of earlier related research.

One example of clearer identification of constructs is in regard to the types of assessment tools used by teachers. This is potentially a complex construct, although on the surface, seems reasonably straightforward. Additional considerations should include the level of cognitive demands of the assessment tasks, whether the tasks are being used for formative or summative purposes, and whether students were informed of the assessment and its contents and performance expectations. These issues, if appropriately identified and measured, could clarify many of the issues raised throughout the literature.

Estimation of Relationships. Both quantitative and qualitative approaches are important to pursue in the development of a comprehensive understanding of the classroom assessment environment. Experimental methods and quasi-experimental approaches will be necessary to uncover some of the important determinants of achievement and performance related to classroom assessment practices. Longitudinal approaches will also provide additional causal evidence and secure a clearer understanding of potentially nonrecursive relationships--causal effects that go both ways.

Finally, the nested nature of students and classrooms must be retained and capitalized in all analyses. Where possible, a third level including school-level characteristics may provide an important organizational component and lead to further understanding of a predominant "culture" or tradition of assessment practices within

schools. Because of the potential influence school administrators or curricular-assessment leaders may provide in local, state, and national reform efforts, school-level characteristics may play an increasingly important role. Improved HLM models will continue to provide appropriate estimation techniques, particularly if sufficient numbers of classrooms nested within schools are included in the analyses--requiring a three-level HLM model.

As will be discussed below, the role of elements of qualitative research should be given serious consideration. Many of the interesting quantitative relationships could be more appropriately interpreted given a more in-depth rich description of classroom assessment process. Quantitative estimation is as much story-telling as qualitative investigation. Combined, the two have great potential for strengthening any line of research.

Inference. The role of validity in this line of research is important at several levels, including construct measurement, study design, estimation, and interpretation. The appropriateness of the inferences made based on the results of any study must be evaluated. This implies a thorough evaluation of the fit between the model and the data gathered, which, as mentioned above, is often taken too lightly or ignored. Perhaps one reason why this is so is because of the technical nature of testing the fit of a model to the data. Testing fit is clearly important because if the model does not adequately fit the data, then the interpretation of estimated parameters is based more on supposition than any model representation of the event under investigation.

Attention to the appropriateness of inferences also implies a role for including the subjects of the investigation in the interpretation of results. Interpretive validity

(Maxwell, 1992) is a concept purported by some qualitative researchers. Interpretive validity suggests that validity is achieved in the interpretation of events when that interpretation is acceptable to the actors involved in the account (e.g., teachers and students). This is a challenge that should be adopted by quantitative researchers involved in theory construction regarding such complex environments as classrooms. Students and teachers could potentially contribute a great deal to the interpretation of results, particularly in regard to how they see things working within their own classrooms.

Similarly, generalizability should also play a critical role in future research. The strength of any resulting theory regarding the classroom assessment environment will derive from the evaluation of diverse classrooms. Based on the complex results of this study, the effects of classroom assessment practices are contextual and interact in unique ways with student characteristics. Future work will need to be done to evaluate these relationships at various grade levels and in various school subjects.

Adjunct Research Tasks

The role of the teacher's implicit and explicit theories of learning and their assessment practices is an area that is potentially important, but not directly related to this research agenda. Little is known at this point regarding the role that a teacher's beliefs about learning plays in their use of assessment tasks (Koellner, Bote, & Middleton, 1998; Phillips & Soltis, 1991).

Although several researchers have attempted to describe teachers' practices, assess teachers' competence in measurement, and evaluate the quality of some assessment tools used by teachers, little is known about the role of training and development activities in

providing the assessment related skills teachers have or do not have. To secure strong evidence of the role of training, these relationships must be understood.

Homework is an area that proved interesting in this study and one which has resulted in inconsistent findings in earlier studies (Cooper, 1989). Homework is an area that teachers, parents, administrators, and students know well -- all having their own opinions on what is optimal. The use of homework did not appear to have any relationship with achievement or other student characteristics except for the frequency with which it was assigned. In fact, out of several types of homework activities used by teachers in the TIMSS data, none had an independent relationship with achievement, although frequency of homework assigned was partially confounded with frequent use of textbook problem sets. Similarly with assessment tools, the labels and definitions used to describe homework activities and uses must be clarified. In some cases, it has not been clear as to whether or not homework included work completed in school and whether or not it involved help from someone outside of school. These considerations could change the way the role of homework is interpreted and used to inform policy.

Policy Implications

This work was driven, in part, by the call for accountability and reform in state and national policies regarding public education. As a vehicle for accountability and reform, assessment plays a central role in these efforts (Johnston & Sandham, 1999; Sack, 1999). As argued earlier, many of the policies which rely on the role of assessment have been forged with little evidence regarding the role of assessment or its capacity to fulfill the expectations of policy makers, as a tool of accountability and reform. Although

this study has not completely delineated the role of assessment practices in motivating student effort and achievement, it has demonstrated the strength of the relationships that exist between classroom assessment practices, student characteristics, and achievement (in a low-stakes settings. With further work on issues raised above, evidence may become available to better inform education policy at all levels.

At this time, an argument can be made for more formally establishing the need for high quality assessment practices at all levels (i.e., classroom, school, district, state, and national). Because of the complex nature of the interactions between classroom assessment practices and student characteristics in their relationships to achievement and because of the focus on the use of assessment for accountability and reform, the following implications are offered for policies related to educational assessment.

Teacher Preparation and Professional Development. Measurement specialists have argued repeatedly that educational measurement is an important part of teacher preparation and that most teachers are poorly prepared to deal with the measurement tasks they face daily (these arguments were reviewed above). The complex nature of the classroom requires a complex understanding of the role and potential impact of assessment. These elements include, but are not limited to, clarifying the objectives of the course given the content demands, considering the pre-requisite skills of the students, considering the content related self-efficacy and attributional styles of students, and considering the appropriateness, consequences, as well as costs and benefits, of various assessment practices.

Teacher preparation programs must begin to adopt the role of preparing teachers for the accountability driven policies they will face in their schools, state, and the nation.

Part of that preparation must include an appreciation for the complexity of the assessment environment as well as provide teachers with the tools and critical decisions skills needed to navigate that environment. The achievement of all students will likely depend on the congruence of all teachers' instructional and assessment activities.

Additionally, teachers need to be able to communicate their goals and values regarding the course content and important skills to facilitate the likelihood of their assessment practices to enhance learning. Although the role of feedback was unidentified in this study, future research should continually uncover this potentially important factor.

If teachers are to continually develop their skills in these areas, some professional development activities must also be devoted to these issues. As states develop accountability systems with state-testing programs as their cornerstone, teachers (as well as school districts as a whole) must be able to make educationally relevant connections between their own assessment practices and the state's assessment system. Professional development opportunities are important tools in this effort.

Administrator Assessment Competence. Trevisan (1999) argued that administrators deficient in their skills regarding assessment are unlikely able to meet their professional responsibilities. This is an important consideration to the extent that school administrators are in substantial positions to support the classroom assessment practices and activities of their teachers. They can also play critical roles in providing connections between classroom assessment activities and those of the district and state. They can develop and promulgate district assessment policies and should be able to communicate assessment results effectively. However, "no state requires assessment competencies advocated by the American Association of School Administrators, National Association

of Elementary School Principals, National Association of Secondary School Principals, and the National Council on Measurement in Education" (Trevisan, p.1).

If teacher preparation programs begin to improve the assessment related training of teachers, administrator preparation programs should do the same. If states adopt assessment competency requirements for certification or licensing of teachers, they should do the same for administrator certification.

School Assessment Policies. Stiggins (1994) provided a sample school district assessment policy that addressed the philosophical base, focus, roles, and responsibilities for all staff engaged in assessment activities in the district. Aside from standards being promoted by content related professional organizations and more generally by measurement and educational professional organizations, few efforts are in place to institutionalize a commitment to responsible use of assessments. Exclusions exist in areas such as special education, where use of assessment has been legislated or mandated through case law. However, the development of a school district assessment policy has the potential to promote responsible use and secure the other implications mentioned here.

State Accountability Systems. "Accountability programs that combine state-adopted academic standards, mandated tests, and related systems of rewards and penalties have been the states' most powerful lever for change" (Johnston & Sandham, 1999, p.19); only Iowa and Nebraska do not have state standards. The role and impact of statewide testing programs is still unclear (Mehrens, 1998). Given the evidence from this project regarding the complex nature of the assessment environment, state policies have far outrun the availability of evidence to support their aims (and, subsequently, their claims).

States have apparently put great efforts into the development of curriculum standards, performance expectations, and in some cases, instructional standards. However, relatively little has occurred with respect to assessment standards. Although a few field-specific professional organizations (e.g., NCTM) have developed assessment standards, these have not been adopted, as have their content standards. The design of statewide testing programs as a component of an accountability system must explicitly recognize the complex interactions between assessment practices and student characteristics. Until more evidence is obtained delineating the connections between statewide assessment outcomes and achievement, and schools are able to measure real gains made by their students, consequences of accountability systems, good or bad, will remain illusive.

Final Thoughts

Much of the motivation to develop and pursue this line of research comes from the questionable results of poor educational research that make headline news (much like that reported in the introduction) and for poorly informed public policy. The complex nature of classrooms should be investigated with approaches equal to the task. Results should be evaluated with equal effort and accompanied by an evaluation of the inferences made. Only then will those in this field be able to achieve gains in understanding such complex environments. Even though the "unruly error terms" that accompany the estimation of any relationship involving "willful human action" may remain large, as Labaree submitted, the importance of uncovering any of the conditions and contexts through which performance gains are made is nonetheless great--great enough to pursue with serious effort.

BIBLIOGRAPHY

Airasian, P. W. (1994). *Classroom assessment* (2nd ed.). New York: McGraw-Hill, Inc.

Airasian, P. W., & Jones, A. M. (1993). The teacher as applied measurer: Realities of classroom measurement and assessment. *Applied Measurement in Education, 6*(3), 241-254.

AFT, NCME, & NEA (1990). *Standards for Teacher Competence in Educational Assessment of Students.* Washington, DC: Authors.

Ames, C., & Archer, J. (1988). Achievement goals in the classroom: Students' learning strategies and motivation processes. *Journal of Educational Psychology, 80*(3), 260-267.

Anderson, B., Mead, N., & Sullivan, S. (1986). *Homework: What do national assessment restuls tell us?* Princeton, NJ: Educational Testing Service.

Beaton, A. E., Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Kelly, D. L., & Smith, T. A. (1996). *Mathematics achievement in the middle school years: IEA's third international mathematics and science study (TIMSS).* Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Bielinski, J., & Davison, M. L. (1998). Gender differences by item difficulty interactions in multiple-choice mathematics items. *American Educational Research Journal, 35*(3), 455-476.

Blumenfeld, P. C., Puro, P., & Mergendoller, J. R. (1992). Classroom Learning and Motivation: Clarifying and Expanding Goal Theory. *Journal of Educational Psychology, 84*(3), 272-281.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: John Wiley and Sons.

Brookhart, S. M. (1994). Teachers' grading: Practice and theory. *Applied Measurement in Education, 7*(4), 279-301.

Brookhart, S. M. (1997). A theoretical framework for the role of classroom assessment in motivating student effort and achievement. *Applied Measurement in Education, 10*(2), 161-180.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models.* Newbury Park, NJ: Sage Publications, Inc.

Bryk, A. S., Raudenbush, S. W., & Congdon, R. (1996a). *HLM* (Version 4.03) [Computer software]. Chicago: Scientific Software International.

Bryk, A. S., Raudenbush, S. W., & Congdon, R. (1996b). *HLM: Hierarchical linear and nonlinear modeling with the HLM/2L and HLM/3L programs*. Chicago: Scientific Software International.

Camp, R. (1992). Assessment in the context of schools and school change. In H. Marshal (Ed.), *Redefining student learning: Roots of educational change*. Grand Forks, ND: University of North Dakota.

Campbell, J. R., Reese, C. M., Sullivan, C. O., & Dossey, J. A. (1996). *NAEP 1994 trends in academic progress*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Carey, L. M. (1994). *Measurement and evaluating school learning* (2nd ed.). Needham Heights, MA: Allyn & Bacon.

Chase, C. I. (1999). *Contemporary assessment for educators*. New York: Longman.

Cizek, G. J. (1996). Grades: The final frontier in assessment reform. *NASSP-Bulletin, 80*(584), 103-110.

Cizek, G. J. (1997). Learning, achievement, and assessment: Constructs at a crossroads. In G. D. Phye (Ed.), *Handbook of classroom assessment: Learning, achievement and adjustment* (pp. 2-32). San Diego, CA: Academic Press.

Cook, T. D., & Campbell, D. T. (1979*). Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton Mifflin Company.

Cooper, H. (1989). *Homework*. White Plains, NY: Longman.

Cooper, H., Linsay, J. J., Nye, B., & Greathouse, S. (1998). Relationships among attitudes about homework, amount of homework assigned and completed, and student achievement. *Journal of Educational Psychology, 90*(1), 70-83.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58*(4), 438-481.

Cross, L. H., & Frary, R. B. (1999). Hodgepodge grading: Endorsed by students and teachers alike. *Applied Measurement in Education, 12*(1), 53-72.

Crowly, M. (1997). Aligning assessment with classroom practices: A promising format. *The Mathematics Teacher, 90*(9), 706-711.

Cunningham, G. K. (1998). *Assessment in the classroom: Constructing and interpreting texts*. Washington, DC: The Falmer Press.

Davenport, E. C., Davison, M. L., Kuang, H., Ding, S. Kim, S-K., & Kwak, N. (1998). High school mathematics course-taking by gender and ethnicity. *American Educational Research Journal, 35*(3), 497-514.

Dempster, F. N. (1997). Using tests to promote classroom learning. In R. F. Dillon (Ed.), *Handbook on testing* (pp. 332-346). Westport, CT: Greenwood Press.

Ebel, R. L. (1976). Measurement and the teacher. In W. A. Mehrens (Ed.), *Readings in Measurement and Evaluation in Education and Psychology* (pp. 73-80). New York: Holt, Rinehart, and Winston.

Ebel, R. L. (1982). Proposed solutions to two problems of test construction. *Journal of Educational Measurement, 19*(4), 267-278.

Ebel, R. L., & Frisbie, D. A. (1986). *Essentials of educational measurement* (4th ed.). Englewood Cliffs, NJ: Printice-Hall, Inc.

Eccles, J., & Midgley (1989). In C. Ames & R. Ames (Eds.), *Research on motivation in education: Goals and cognitions* (Vol. 3, pp. 283-331).

Edwards, V. B. (1998, January 8). Quality Counts '98: An Education Week/Pew Charitable Trusts report on education in the 50 states. *Education Week, 17*(17), p. 6.

Frank, K. A. (1999). Quantitative methods for studying social context in multilevels and through interpersonal relations. *Review of Research in Education, 23*, 171-216.

Gallagher, J. D. (1998). *Classroom measurement for teachers*. Upper Saddle River, NJ: Prentice-Hall, Inc.

Gitomer, D. H., & Duschl, R. A. (1994). *Moving toward a portfolio culture in science education* (MS No. 94-07). Princeton, NJ: Educational Testing Service.

Glaser, R. & Silver, E. (1994). Assessment, testing, and instruction: Retrospect and prospect. *Review of Research in Education, 20*, 393-419.

Glasser, W. (1985). *Control theory in the classroom*. New York: Harper & Row, Publishers.

Gonzales, E. J., & Smith, T. A. (Eds.). (1997). *Users guide for the TIMSS International database*. Amsterdam, The Netherlands: IEA. Available: http:// www.csteep.bc.edu/ timss1/database/ UG_1and2.pdf

Good, T. L., & Brophy, G. E. (1986). *Educational psychology* (3rd ed.). New York: Longman.

Gredler, M. E. (1999). *Classroom assessment and learning*. New York: Longman.

Gronlund, N. E. (1985). *Measurement and Evaluation in Teaching* . New York: Macmillan.

Haladyna, T. M. (1994). *Developing and validating multiple-choice test items*. Hillsdale, NJ: Lawrence Earlbaum Associates, Publishers.

Haladyna, T. M. (1997). *Writing test items to evaluate higher order thinking*. Boston: Allyn and Bacon.

Hanna, G. S. (1993). *Better teaching through better measurement*. Orlando, FL: Harcourt Brace Jovanovich Publishers.

Hopkins, K (1998). *Educational and psychological measurement* (8th ed.). Needham Heights, MA: Allyn & Bacon.

Johnston, R. C., & Sandham, J. L. (1999, April 14). States increasingly flexing their policy muscle. *Education Week, 18*(31), pp. 1, 19-20.

Jöreskog, K. G., & Sörbom, D. (1998). *LISREL* (Version 8.20) [Computer software]. Chicago: Scientific Software International, Inc.

Kachigan, S. K. (1991). *Multivariate statistical analysis: A conceptual introduction*. New York: Radius Press.

Kasim, R. M., & Raudenbush, S. W. (1998). Application of Gibbs sampling to nested variance components with heterogeneous within-group variance. *Journal of Educational and Behavioral Statistics, 23*(2), 93-116.

Keith, T. Z., & Cool, V. A. (1992). Testing models of school learning: Effects of quality of instruction, motivation, academic coursework, and homework on academic achievement. *School Psychology Quarterly, 7*(3), 207-226.

Keith, T. Z., Keith, P. B., Troutman, G. C., Bickley, P. G., Trivette, P. S., & Singh, K. (1993). Does parental involvement affect eighth-grade student achievement? *School Psychology Review, 22*, 474-496.

Koellner, K. A., Bote, L. A., & Middleton, J. A. (1998, April). *Cycles of transformation in assessment practices in a cognitively guided instruction classroom*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Kreft, I., & De Leeuw, J. (1998). *Introducing multilevel modeling*. London: Sage Publications.

Kubiszyn, T., & Borich, G. (1996). *Educational testing and measurement* (3rd ed.). Lenview, IL: Scott, Foresman and Company.

Labaree, D. F. (1998). Educational researchers: Living with a lesser form of knowledge. *Educational Researcher, 27*(8), 4-12.

Lindquist, E. F. (1936). The theory of test construction. In H. E. Hawkes, E. F. Lindquist, and C. Mann (Eds.), *The construction and use of achievement examinations*. Boston: Houghton Mifflin Company.

Linn, R. L., & Gronlund, N. (1995). *Measurement and assessment in teaching* (7th ed.). Columbus, OH: Merrill.

Marsh, H. W., & Craven, R. (1997). Academic self-concept: Beyond the dustbowl. In G. D. Phye (Ed.), *Handbook of classroom assessment: Learning, achievement and adjustment* (pp. 131-198). San Diego, CA: Academic Press.

Maxwell, J. A. (1992). Understanding and validity in qualitative research. *Harvard Educational Review, 62*(3), 279-300.

McDaniel, E. (1994). *Understanding educational measurement.* Madison, WI: Brown & Benchmark Publishers.

McMillan, J. H. (1997). *Classroom assessment: Principles and practice for effective instruction.* Needham Heights, MA: Allyn & Bacon.

McMorris, R. F., & Boothroyd, R. A. (1993). Tests that teachers build: An analysis of classroom tests in science and mathematics. *Applied Measurement in Education, 6*(4), 321-342.

Mehrens, W. A. (1998). Consequences of assessment: What is the evidence? *Education Policy Analysis Archives, 6*(13).

Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and Evaluation in Education and Psychology.* New York: Holt, Rinehart & Winston.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5-11.

Miller, N., & Pollock, V. E. (1994). Meta-analytic synthesis for theory development. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 457-483). New York: Russell Sage Foundation.

Natriello, G. (1987). The impact of evaluation processes on students. *Educational Psychologist, 22*, 155-175.

NCTM (1998). *Principles and Standards for School Mathematics: Electronic Version.* Available: http://standards-e.nctm.org/1.0/normal/standards/frnt_MAIN.html

Newmann, F. M. (1997). Authentic assessment in social studies: standards and examples. In G. D. Phye (Ed.), *Handbook of classroom assessment: Learning, achievement and adjustment* (pp. 360-380). San Diego, CA: Academic Press.

Nitko, A. J. (1996). *Educational assessment of students* (2nd ed.). Columbus, OH: Merrill.

Olson, L., & Jerald, C. D. (1998a, January 8). Barriers to success. *Education Week 17*(19), p. 9.

Olson, L., & Jerald, C. D. (1998b, January 8). The achievement gap. *Education Week 17*(17), pp. 10-13.

Oosterhof, A. C. (1990). *Classroom applications of educational measurement.* Columbus, OH: Merrill Publishing Company.

Oosterhof, A. C. (1999). *Developing and using classroom assessments* (2nd ed.). Upper Saddle River, NJ: Prentice-Hall, Inc.

Payne, D. A. (1997). *Measuring and evaluating educational outcomes.* New York: Merrill.

Phillips, D. C., & Soltis, J. F. (1991). *Perspectives on learning* (2nd ed.). New York: Teachers College Press.

Plake, B. S. (1993). Teacher assessment literacy: Teachers' competencies in the educational assessment of students. *Mid-Western Educational Researcher, 6*(1), 21-27.

Plake, B. S., & Impara, J. C. (1997). Teacher assessment literacy: What do teachers know about assessment? In G. D. Phye (Ed.), *Handbook of classroom assessment: Learning, achievement and adjustment* (pp. 55-70). San Diego, CA: Academic Press.

Plake, B. S., Impara, J. C., & Fager, J. J. (1993). Assessment competencies of teachers: A national survey. *Educational Measurement: Issues and Practice*, 10-12.

Popham, W. J. (1990). *Modern educational measurement: A practitioner's perspective* (2nd ed.). Englwood Cliffs, NJ: Prentice-Hall, Inc.

Reynolds, P. D. (1971). *Primer in theory construction.* Indianapolis, IN: The Bobbs-Merril Company, Inc.

Rodriguez, M. C. (1995). *Minnesota higher education: Modeling the state policy environment.* Unpublished masters thesis. Minneapolis, MN: University of Minnesota.

Sack, J. L. (1999, January 27). Clinton links K-12 dollars, performance. *Education Week, 18*(20), pp. 1, 21.

Salmon-Cox, L. (1980, April). *Teachers and tests: What's really happening?* A paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Sax, G. (1997). *Principles of educational and psychological measurement and evaluation* (4th ed.). New York: Wadsworth Publishing Company.

Schmidt, W. H. (1993, May). *TIMSS: Concepts, measurements and analyses, Survey of Mathematics and Science Opportunities*, Research Report Series No. 56. A paper

presented at the annual meeting of the American Educational Research Association, Los Angeles, CA.

Schmidt, W. H., McKnight, C. C., & Raizen, S. A. (1996). *A splintered vision: An investigation of U.S. science and mathematics education.* Boston: Kluwer Academic Publishers.

SciMathMN (1996). *Minnesota TIMSS report: A preliminary summary of results.* St. Paul, MN: Author.

Shavelson, R. J., & Stern, P. (1981). Research on teachers pedagogical thoughts, judgments, decisions and behavior. *Review of Educational Research, 51*(4), 455-498.

Shields, P. M., Corcoran, T. B., & Zucker, A. A. (1995). *Evaluation of the National Science Foundation's statewide systemic initiatives program: First year report* (NSF publication no. 94-95). Washington, DC: National Science Foundation.

Stenmark, J. K. (1991, Ed.). *Mathematics assessment: Myths, models, good questions, and practical suggestions.* Reston, VA: National Council of Teachers of Mathematics.

Stiggins, R. J. (1989). Measuring thinking skills through classroom assessment. *Journal of Educational Measurement, 26*(3), 233-246.

Stiggins, R. J. (1991a). Assessment literacy. *Phi Delta Kappan, 72*(7), 534-539.

Stiggins, R. J. (1991b). Relevant classroom assessment training for teachers. *Educational Measurement: Issues and Practice, 10*(1), 7-12.

Stiggins, R. J. (1991c). Facing the challenges of a new era of educational assessment. *Applied Measurement in Education, 4*(4), 263-273.

Stiggins, R. J. (1993). Teacher training in assessment: Overcoming the neglect. In S. L. Wise (Ed.), *Teacher training in measurement and assessment skills* (pp. 27-40). Lincoln, NE: Buros Institute of Mental Measurements.

Stiggins, R. J. (1994). *Student-centered classroom assessment.* New York: Macmillan College Publishing Company.

Stiggins, R. J. (1995a). Professional development: The key to a total quality assessment environment. *NASSP Bulletin, 79*(573), 11-19.

Stiggins, R. J. (1995b). Assessment literacy for the 21st century. *Phi Delta Kappan, 77*(3), 238-245.

Stiggins, R. J. (1997). *Student centered assessment* (2nd ed.). Upper Saddle River, NJ: Merrill.

Stiggins, R. J. (1998). *Learning team training guide & classroom assessment self-evaluation*. Portland, OR: Assessment Training Institute.

Stiggins, R. J. & Bridgeford, N. J. (1985). The ecology of classroom assessment. *Journal of Educational Measurement, 22*(4), 271-286.

Stiggins, R. J., & Conklin, N. F. (1992). *In teachers' hands: Investigating the practices of classroom assessment*. Albany, NY: State University of New York Press.

Stroud, J. B. (1946). *Psychology in education*. New York: David McKay Co., Inc.

Thissen, D. (1991a). *MULTILOG* (Version 6.30) [Computer software]. Chicago: Scientific Software International.

Thissen, D. (1991b). *MULTILOG user's guide*. Chicago: Scientific Software International

Thompson, D. R., Beckmann, C. E., & Senk, S. L. (1997). Improving classroom tests as a means of improving assessment. *The Mathematics Teacher, 90*(1), 58-64.

Thorndike, R. M. (1997). *Measurement and evaluation in psychology and education* (6th ed.). Columbus, OH: Merrill.

Tiegs, E. W. (1931). *Tests and measurements for teachers*. Cambridge, MA: The Riverside Press.

Tindal, G. A., & Marston, D. B. (1990). *Classroom-based assessment: evaluating instructional outcomes*. Columbus, OH: Merrill Publishing Company.

Traug, A. L., & Friedman, S. J. (1996, April). *Evaluating high school teachers' written grading policies*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Trevisan, M. S. (1999). Administrator certification requirements for student assessment competence. *Applied Measurement in Education, 12*(1), 1-11.

Van Moorlehem, T. (1998a, January 19). Home sales, custody fights hinge on exam. *Detroit Free Press*.

Van Moorlehem, T. (1998b, April 29). Students, parents rebel against state test. *Detroit Free Press*, pp. A1, A8.

Van Moorlehem, T., & Newman, H. (1998, January 19). Testing MEAP scores: Comparing raw results has some pitfalls. *Detroit Free Press*.

Walberg, H. J. (1991). Does homework help? *School Community Journal, 1*(1), 13-15.

Ward, A. W., & Murray-Ward, M. (1999). *Assessment in the classroom*. Belmont, CA: Wadsworth Publishing Company.

Weirsma, W., & Jurs, S. G. (1990). *Educational measurement and testing* (2nd ed.). Boston: Allyn & Bacon.

Wigfield, A., Eccles, J. S., & Rodriguez, D. (1999). The development of children's motivation in school contexts. *Review of Research in Education, 23*, 73-118.

Wiggins, G. P. (1999). *Educative assessment: Designing assessments to inform and improve student performance*. San Francisco, CA: Jossey-Bass Publishers.

Worthen, B. R., Borg, W. R., & White, K. R. (1993). *Measurement and evaluation in the schools*. New York: Longman Publishing Group.

Worthen, B. R., White, K. R., Fan. X., & Sudweeks, R. R. (1999). *Measurement and assessment in schools* (2nd ed.). New York: Addison Wesley Longman, Inc.

Zhang, Z. (1996, April). *Teacher assessment competency: A Rasch model analysis*. Paper presented at the Annual Meeting of the American Educational Research Association, New York.

Zucker, A. A., Shields, P. M., Adelman, N., & Powell, J. (1995). *Evaluation of the National Science Foundation's statewide systemic initiatives program: Second year report* (NSF publication no. 95-64). Washington, DC: National Science Foundation.

APPENDIX A

This appendix contains tables for percent responding to each option for each of the variables reported above. It also contains companion tables for science teachers and students' science related attitudes and behaviors. These were provided for comparison purposes. The tables on teachers included 326 mathematics teachers and 294 science teachers (including those science teachers who had completed background questionnaires with at least six students in the student database).

The tables on students included the 6963 students with corresponding math teachers (the same set of students used in the above analyses for mathematics classrooms was used to report science attitudes and behaviors as well). Student percentages were weighted by their probability for selection in the sample.

Table A- 1

How often do you usually assign homework?

	Never	Less than once a week	Once or twice a week	3 or 4 times a week	Every day
Mathematics	1	3	12	57	27
Science	2	13	37	36	8

Table A- 2

If you assign homework, how many minutes of homework do you usually assign your students?

	Do not assign homework	Less than 15 minutes	15-30 minutes	31-60 minutes	60-90 minutes
Mathematics	1	8	70	19	1
Science	4	15	66	12	0

Table A- 3

If you assign mathematics homework, how often do you assign each of the following kinds of tasks?

	Do not assign homework	Never	Rarely	Sometimes	Always
Problem/question sets in textbooks	1	4	7	60	28
Worksheets or workbook	1	5	18	69	8
Finding one or more uses of the content covered	1	24	32	38	6
Small investigations or gathering data	1	21	37	41	1
Reading in textbook or supplementary material	1	32	32	31	4
Writing definitions or other short assignments	1	34	38	26	2
Working individually on long term projects	1	34	36	28	1
Working as small group on long term projects	1	42	36	20	0
Preparing oral reports, individual or group	1	50	34	15	1
Keeping a journal	1	59	21	12	7

Table A- 4

If you assign science homework, how often do you assign each of the following kinds of tasks?

	Do not assign homework	Never	Rarely	Sometimes	Always
Problem/question sets in textbooks	2	10	16	61	11
Worksheets or workbook	2	10	18	63	7
Reading in textbook or supplementary material	2	13	19	54	12
Writing definitions or other short assignments	2	9	24	57	8
Small investigations or gathering data	2	6	30	58	3
Working individually on long term projects	2	7	32	53	6
Preparing oral reports, individual or group	2	18	31	46	3
Finding one or more uses of the content covered	2	17	42	34	5
Working as small group on long term projects	2	23	35	38	2
Keeping a journal	2	46	22	21	9

Table A- 5

If students are assigned written mathematics homework, how often do you do the following?

	Do not assign homework	Never	Rarely	Sometimes	Always
Record whether or not homework was completed	1	1	1	17	80
Use it to contribute towards grades	1	1	4	28	67
Give feedback to whole class	1	1	4	37	57
Collect, correct, and return to students	1	6	15	45	34
Use it as a basis for class discussion	1	2	10	63	25
Have students correct own work in class	1	4	11	52	32
Have students exchange assignments and correct in class	1	25	22	43	8
Collect, correct, and keep assignments	1	36	24	31	9

Table A- 6

If students are assigned written science homework, how often do you do the following?

	Do not assign homework	Never	Rarely	Sometimes	Always
Record whether or not homework was completed	2	1	0	15	80
Use it to contribute towards grades	2	1	3	31	64
Give feedback to whole class	2	1	5	40	52
Collect, correct, and return to students	2	2	9	38	48
Use it as a basis for class discussion	2	2	8	69	19
Have students correct own work in class	2	12	21	56	8
Have students exchange assignments and correct in class	2	20	22	52	3
Collect, correct, and keep assignments	2	31	19	33	15

Table A- 7

In assessing the work of the students in your mathematics class, how much weight do you give each of the following types of assessment?

	None	Little	Quite a Lot	A Great Deal
Teacher-made short answer or essay tests requiring students to explain their reasoning	13	34	37	15
How well students do on homework assignments	5	44	47	5
Observations of students	14	43	36	7
Responses of students in class	15	41	36	8
How well students do on projects or practical/laboratory exercises	28	37	31	3
Teacher-made multiple-choice, true-false and matching tests	32	44	21	4
Standardized tests produced outside the school	36	44	19	1

Table A- 8

In assessing the work of the students in your science class, how much weight do you give each of the following types of assessment?

	None	Little	Quite a Lot	A Great Deal
How well students do on projects or practical/laboratory exercises	2	20	63	15
Teacher-made short answer or essay tests requiring students to explain their reasoning	4	32	49	15
Teacher-made multiple-choice, true-false and matching tests	7	24	58	11
How well students do on homework assignments	4	40	50	6
Observations of students	9	42	44	5
Responses of students in class	12	46	38	3
Standardized tests produced outside the school	45	41	12	1

Table A- 9

How often do you use the mathematics assessment information you gather from students to...

	None	Little	Quite a Lot	A Great Deal
Provide students' grades	1	5	57	37
Provide feedback to students	0	7	67	25
Plan for future lessons	0	14	62	23
Report to parents	1	20	61	18
Diagnose students' learning problems	2	21	61	17
Assign students to different programs or tracks	20	47	26	7

Table A- 10

How often do you use the science assessment information you gather from students to...

	None	Little	Quite a Lot	A Great Deal
Provide students' grades	1	6	57	37
Provide feedback to students	1	10	69	20
Plan for future lessons	3	23	58	17
Report to parents	1	27	55	17
Diagnose students' learning problems	3	40	46	11
Assign students to different programs or tracks	40	39	17	3

Table A- 11

To do well in mathematics at school, you need...

	Strongly Agree	Agree	Disagree	Strongly Disagree
Natural Talent	15	35	43	7
Good Luck	11	21	50	17
Lots of hard work studying	53	37	8	2
To memorize notes	20	38	32	10

Table A- 12

To do well in science at school, you need...

	Strongly Agree	Agree	Disagree	Strongly Disagree
Natural Talent	17	34	42	8
Good Luck	13	22	48	17
Lots of hard work studying	52	38	8	2
To memorize notes	25	40	26	8

Table A- 13

What do you think about mathematics?

	Strongly Agree	Agree	Disagree	Strongly Disagree
I enjoy learning mathematics	21	50	21	7
Mathematics is boring	16	28	41	15
Mathematics is an easy subject	15	33	37	15
Mathematics is important to everyone's life	57	35	5	3
I would like a job that involved using mathematics	15	34	32	20

Table A- 14

What do you think about science?

	Strongly Agree	Agree	Disagree	Strongly Disagree
I enjoy learning science	25	50	17	8
Science is boring	13	25	44	18
Science is an easy subject	15	40	35	10
Science is important to everyone's life	31	48	16	4
I would like a job that involved using science	20	29	31	19

Table A- 15

How well do you usually do in mathematics and science at school?

	Strongly Agree	Agree	Disagree	Strongly Disagree
I usually do well in mathematics	37	49	11	2
I usually do well in science	36	51	10	2

Table A- 16

How much do you like...

	Dislike a lot	Dislike	Like	Like a lot
Mathematics?	12	16	48	25
Science?	11	15	47	26

Table A- 17

My mother thinks it is important for me to...

	Strongly agree	Agree	Disagree	Strongly disagree
Do well in mathematics at school	72	26	1	1
Do well in science at school	64	34	2	1

APPENDIX B

This appendix contains the descriptions of each topic taught by the mathematics teachers, as they were presented in the Teacher Background Questionnaire. The entire set of Questionnaires is available at the following web-site:

http://www.csteep.bc.edu/timss

IEA (1994). Teacher Questionnaire (Mathematics) Population 2. Chestnut Hill, MA: TIMSS Study Center, Boston College.

How long did you spend teaching each of these topics in your math class this year?

a) Whole Numbers
   1. Meaning of whole numbers; place value and numeration
   2. Operations with and properties of whole numbers

b) Common & Decimal Fractions
   1. Meaning, Representation and Uses of Common Fractions
   2. Properties of Common Fractions
   3. Meaning, Representation and Uses of Decimal Fractions
   4. Properties of Decimal Fractions
   5. Relationships Between Common and Decimal Fractions
   6. Conversion of Equivalent Forms
   7. Ordering of Fractions (Common and Decimals)

c) Percentages
   Concepts of percentage; computation with percentage; types of percentage problems

d) Number Sets & Concepts
   Uses, properties, and computations with integers (negative as well as positive), rational numbers (including negative fractions), real numbers complex numbers; number bases other than ten; exponents, roots and radicals.

e) Number Theory
   Prime and composite numbers; factorizations of whole numbers; greatest common divisors; least common multiples; permutations; combinations; systematic counting of possibilities and so on

f) Estimation & Number Sense
   Estimating quantity and size; rounding and significant figures, estimating the results of computations (including mental arithmetic and reasonableness of results); scientific notation and orders of magnitude

g) Measurement Units & Processes
   Ideas and units of measurement; standard metric units; length, area, volume, capacity, time, money and so on; use of measurement instruments

h) Estimation & Error of Measurement
   Estimation of measurements other than perimeter and area; precision and accuracy; errors of measurement

i) Perimeter, Area, & Volume
   Perimeter & area of trianges, quadrilaterals, polygons, circles & other two-dimensional shapes; Calculating, estimating, & solving problems involving perimeters and areas; Surface area and volume

j) Basics of One & Two Dimensional Geometry
Number lines and graphs in one and two dimensions; triangles, quadrilaterals, other polygons, and circles; equations of straight lines; Pythagorean Theorem

k) Geometric Congruence & Similarity
Concepts, properties and uses of congruent and similar figures, especially for triangles, quadrilaterals, other polygons and plan shapes

l) Geometric Transformations & Symmetry
Geometric patterns; tessellations; kinds of symmetry in geometric figures, symmetry of number patterns; transformations of all types and their representations; algebraic structure and properties of sets of transformations

m) Constructions & Three Dimensional Geometry
Constructions with compass and straightedge; conic sections; three-dimensional shapes, surfaces and their properties; lines and planes in space; spatial perception and visualization; coordinate graphs and vectors in three dimensions

n) Ratio & Proportion
1. Concepts and Meaning
2. Applications and Uses
Maps and models; solving practical problems based on proportionality; solving proportional equations

o) Proportionality: Slope, Trigonometry & Interpolation
1. Slope and Trigonometry
Slope; trigonometric ratios; solving triangles and problems involving triangles including the rules of sines and of cosines
2. Linear Interpolation and Extrapolation

p) Functions, Relations, & Patterns
Number patterns; relations, their properties and graphs; types of function (linear, quadratic, exponential, trigonometric, inverse, etc.); operations on functions; relations of functions and equations (roots, zeros, etc.); problems involving functions

q) Equations, Inequalities, & Formulas
1. Linear Equations and Formulas
Representing situations algebraically; work with formulas other than measurement formulas; algebraic expressions & working with them (Factoring, polynomial operations, etc.); solving linear equations
2. Other Equations and Formulas
Solving various types of equations (quadratic, radical, trigonometric, logarithmic, etc.); inequalities; systems of equations; systems of inequalities

r) Statistics & Data
Collecting data from experiments & surveys; representing & interpreting data in tables, charts, graphs, etc; nominal, ordinal, etc., scales; means, medians & other measures of central tendency; variance, standard deviations & other measure of dispersion; sampling, randomness & bias; prediction & inferences from data;

regression & fitting lines & curves to data; correlation's & other measures of relationship; use & misuse of statistics in analyzing data

s) Probability & Uncertainty
Informal language of 'more likely,' 'less likely', etc.; probability models & numerical probability; all other aspects of probability & probability distributions for random variables; expectations, parameter estimation, hypothesis testing, confidence intervals, & related statistical topics

t) Sets & Logic
Sets, set notation and set operations; classification; logic and truth tables

u) Problem Solving Strategies
Problem solving heuristics and strategies

v) Other Mathematics Content
Mark here for all content you covered that was not in one of the earlier categories. This includes advanced topics such as the following: Computers (operation of computers, flow charts, learning a programming language, programs, algorithms with applications to the computer); History and nature of mathematics; and Proofs.