APPLIED MEASUREMENT IN EDUCATION, *17*(1), 1–24

The Role of Classroom Assessment in Student Performance on TIMSS

Michael C. Rodriguez
Department of Educational Psychology
University of Minnesota

This project evaluated the relationship between assessment practices and achievement and the mediating roles of student self-efficacy and effort. In part, this was based on a framework proposed by Brookhart (1997). The United States portion of the Third International Math and Science Study was used to estimate these relationships. Several student level characteristics were important explanatory variables regarding variation in mathematics achievement, including mathematics self-efficacy, effort, and level of uncontrollable attributions. At the classroom level, teacher assessment practices had significant relationships to classroom performance. In addition, cross-level interactions (between student characteristics and teacher practices) suggested that classroom assessment practices might uniquely interact with student characteristics in their role of motivating student effort and performance.

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, and 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 so doing, impact the learning process. 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 aca-

Requests for reprints should be sent to Michael C. Rodriguez, Department of Educational Psychology, University of Minnesota, 206 Burton Hall, 178 Pillsbury Drive SE, Minneapolis, MN 55455-0211. E-mail: mcrdz@umn.edu

demic motivation, and organizing and securing the storage of knowledge and skills (for reviews, see Black & Wiliam, 1998; Crooks, 1988; Dempster, 1997). Assessment at the classroom level is clearly important.

Measurement specialists have suggested improvements in classroom measurement-related professional development. To contribute to this effort, measurement specialists should attempt to communicate with a broader audience concerning the merits of best practice, particularly outside of the measurement journals (Cross & Frary, 1999). We are all aware of the negative consequences of limited measurement knowledge in the practice of assessment.

Requirements for certification, topics of professional development, and standards of practice are not substantially informed by evidence. 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 (a) broaden the scope of coverage in understanding key relationships in the classroom assessment environment, (b) illustrate an analytical method that takes advantage of the nested nature of classroom data, and (c) illuminate critical methodological issues for future research efforts.

One attempt to specify theoretical links between assessment and achievement is a model of the role of classroom assessment in motivating student effort and achievement proposed by Brookhart (1997). The model suggests that the classroom assessment environment "plays out" in repeated assessment events through which a teacher communicates and students respond.

In this study, mathematics was the subject area chosen because of the national focus on mathematics within Goals 2000 and No Child Left Behind, and the availability of data from the comprehensive mathematics assessment used by the Third International Math and Science Study (TIMSS). Middle-school classrooms were chosen because of the importance of the transition period from elementary school to high school, where curricular differentiation is greatest. The primary question to be answered was, "What are the interrelationships of teacher assessment practices, student self-efficacy, student effort, and achievement performance?"

BACKGROUND

"Classroom teachers are the ultimate purveyors of applied measurement, and they rely on measurement and assessment-based processes to help them make decisions every hour of every school day" (Airasian & Jones, 1993, pp. 241–242). Teachers spend at least one third of their professional time on assessment activities that inform a wide variety of decisions made daily and directly influence students' learning experiences (Stiggins & Conklin, 1992).

Classroom Assessment

Much of the literature regarding classroom assessment exists in the form of professional development-related articles and books. Richard Stiggins at the Assessment Training Institute has been a leader in this literature (see, for example, Stiggins, 1989, 1991, 1993, 2001). Early on, the focus of Stiggins' research on classroom assessment was to describe the ecology of the classroom assessment environment. Stiggins and Bridgeford (1985) surveyed 228 teachers from eight districts around the United States and found that use of teacher-made objective tests increased between 2nd and 11th grade. Half of the teachers who used their own objective tests reported to be comfortable with that type of assessment. Math and science teachers were more likely to use objective tests than writing and speech teachers were. Use of published tests, including norm-referenced tests and tests accompanying text books, decreased across grades, but they 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 when asked about their objective tests focused on test improvement.

Some also argue 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" (Stiggins & Conklin, 1992, p. 2). After observing three sixth-grade classrooms for 10 weeks, Stiggins and Conklin reported, "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), in an early review of the literature on assessment practices, found that teachers relied mostly on their own assessment activities for information on student achievement. Observations and classroom work were also important sources of information. He reported the results of a survey of high school teachers regarding sources of information about the achievement of their students, where 40% used their own tests, 30% used interactions with students, 21% relied on homework performance, 6% used observations of students, and 1% used standardized tests.

A profile has 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). Currently, there is a more active research agenda on what occurs around assessment, that is, learning and achievement, often the targets of assessment.

Learning, Achievement, and Assessment

Definitions of learning have changed subtly to exclude any reference to student behavior and center on cognitive change exclusively (Cizek, 1997). Based on the conceptual work of other researchers and his own conditions for an appropriate and meaningful definition of assessment, Cizek (1997) proposed the following definition:

> 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. (p. 10)

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 that impact learning outcomes—cognitive and noncognitive)" (p. 323). Their model was illustrated in a flowchart, in which student effort impacted performance, whereas instructional factors and student inputs affected both student effort and performance. A unique component, compared to Brookhart's model, was the inclusion of the consequences of achievement that derive from performance, in which the consequences subsequently impact both instructional factors and student inputs.

TIMSS Conceptual Model

The data used in this study came from the TIMSS-USA database. The TIMSS designers conceptualized student learning as being influenced by psychological theories of individual differences and motivation, as well as sociological concepts involving family background. A conceptual model was developed to guide instrument design.

> 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)

The role of effort and motivation as a moderator of achievement and performance was key to the Brookhart model as well.

A Testable Framework

Brookhart (1997) made explicit connections between the role of classroom assessment practices in motivating student effort and achievement while integrating the literatures from classroom assessment environments and social-cognitive theories of learning and motivation. A classroom assessment event consists 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. Such an event includes developing a discrete set of objectives and assessments of whether the objectives were met.

"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" (Brookhart, 1997, p. 167). 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" (Brookhart, 1997, p. 173). The amount of invested mental effort includes the nonautomatic 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). Others have focused on the impact of feedback on motivation as well. Teachers' feedback, accountability, and evaluation practices affect students' motivational orientation, whether they are motivated to learn or simply perform (Ames & Archer, 1988; Blumenfeld, Puro, & Mergendoller, 1992).

In a subsequent test of this theoretical framework, Brookhart and DeVoge (1999) evaluated the hypothesized relationships in two third-grade language arts classrooms. Pre- and postsurveys were conducted around four assessment events and four students were interviewed.

Through careful analysis of descriptive statistics, observational records, and interview data, evidence was provided for several conclusions: (a) The role of classroom assessment model held to the extent that the data demonstrated relationships among student perceptions of tasks, their effort, and achievement; (b) previous experience with similar tasks informed student self-efficacy judgments (the importance of the functional significance of feedback); (c) the relationship between self-efficacy and effort is complex; (d) the functional significance of feedback and the type of feedback has theoretical importance; and (e) the importance of goal orientation was evident in the interview results. Unfortunately, many of the correlations were small and some were unexpectedly negative, in part because of the small sample, limited range of classroom assessment environments in this study and a homogenous group of students in one classroom.

The Brookhart (1997) model is an event model, looking at the interaction of classroom practices, student perceptions and effort, and achievement within a given assessment event. The model adopted for this project is considered a generalization of the Brookhart model based on a more general use of the literature. The generalized model is described later.

Toward a Theory of Classroom Assessment

As suggested earlier, 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 framework for the role of classroom assessment, approached 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 teachers to do” (p. 178). A prescriptive theory of test design “would predict which test design would be most appropriate in a particular instructional procedure under given instructional conditions and for specified instructional outcomes” (Nitko, 1989, p. 417).

> 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 ad hoc 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 potential interactions between teacher decision requirements, instructional activities, student characteristics, and elements of classroom assessment practices.

METHODS

Research Design

This study was primarily cross-sectional, examining existing conditions of several student and classroom characteristics and their relationships to student achievement. The unit of observation (subject of this study) was the student. However, teacher and classroom level data were available, making the resulting data hierarchical: Students were nested within classrooms.

A generalized model was evaluated (Figure 1) based on an extensive review of the literature and the frameworks provided by the TIMSS conceptual model, Brookhart (1997), and Ward and Murray-Ward (1999). The model was formally evaluated as a hierarchical linear model, which allowed for appropriate accounting of the nested nature of the data and included additional features such as demographic information of the students.

Participants

The participants of this study included the middle school mathematics teachers who participated in the TIMSS and their students. Teachers who did not complete the background questionnaire or who had fewer than six students from their classes with complete assessments and background questionnaires were excluded. The final database used to fit the Hierarchical Linear Modeling (HLM) model included 328 teachers and 6,963 students.

Of the 6,963 students, 51% were girls, 36% were in seventh grade (mean age = 13.2, *SD* = 0.55), and 64% were in eighth grade (mean age = 14.2, *SD* = 0.51). Most of the students always spoke English at home (87.2%), whereas others sometimes (11.6%) or never spoke English at home (1.2%).

Of the 328 mathematics teachers, 35% taught seventh grade and 65% taught eighth grade. Just over 67% of the mathematics teachers were women. On average, teachers had 21 students in their class also included in the student database (*SD* = 6). For a more complete discussion of the TIMSS database, see Gonzales and Smith (1997).

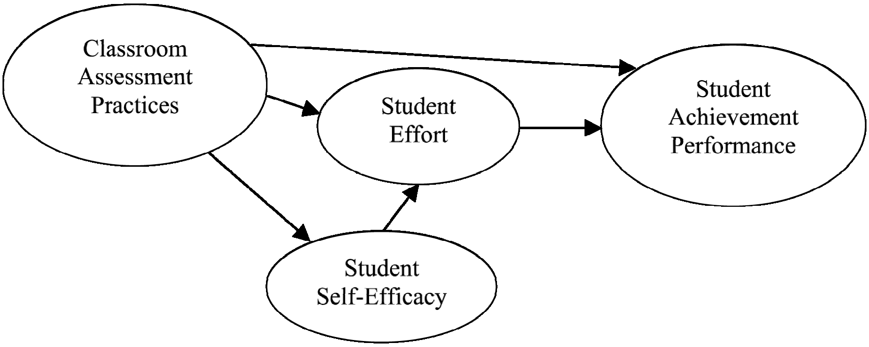
Classroom Assessment Practices
Student Effort
Student Achievement Performance
Student Self-Efficacy

FIGURE 1 The general model evaluated in this study.

TIMSS Instruments

Mathematics assessment instrument. The complete item pool for mathematics consisted of 151 mathematics items (125 multiple choice, 26 constructed response). Each student received a test booklet with a sample of items on forms created through a matrix-sampling process. The overall assessment reliabilities were reported as median alpha coefficients from the eight booklets used in the TIMSS assessment: The mathematics assessment had a reliability of .86 among seventh-grade students and .89 among eighth-grade students. The scores used in this project were two-parameter item response theory (IRT) maximum likelihood estimates of theta (ability) based on items calibrated using Multilog 6.3.

TIMSS background questionnaires. The student questionnaires were administered separately from the assessment instruments, taking between 20 and 40 min. The student questionnaires asked students about their personal backgrounds and home environment, academic activities, 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 and assessment practices, students' opportunity to learn various topics and their pedagogic beliefs.

Student Level Constructs

The primary goal of this project was to uncover the relationships between teacher classroom assessment practices and student achievement performance. 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. The TIMSS database provided no information related to student perceptions of the assessment task.

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. Although three indicators were available regarding possible kinds of feedback students received regarding the results of their homework (correcting the homework of other students, having the teacher correct homework, and discussing completed homework as a class), their intercorrelations were weak with weak corresponding relationships to achievement. Feedback was eliminated from further analyses.

Student self-efficacy. Several items in the student questionnaire addressed mathematics self-efficacy (student perceptions of their potential for mastery of mathematics) and attribution of control in mathematics. Among indicators of mathematics self-efficacy was an item asking students if they agreed (on a 4-point scale) with the following statements: (a) I like mathematics, (b) I enjoy learning mathematics, (c) Math is an easy subject, and (d) I would like a job involving mathematics. A structural equation model was used to confirm the combined items as a measure of what was considered mathematics self-efficacy. The model fit exceptionally well with a root mean square error of approximation (RMSEA) of .05. The RMSEA is seen as an improvement over other residual point estimate indicators of fit because it considers a 90% confidence interval about the model-implied versus observed covariance matrix. Values of RMSEA below .10 are seen as a good fit with .05 or less as very good fit (Steiger, 1990). The structural equation model weights were used to combine the four items into a self-efficacy scale.

Regarding attribution of control, students reported the degree to which they agreed (4-point scale) that to do well in mathematics at school, they need (a) lots of natural talent, (b) good luck, (c) lots of hard work studying at home, and (d) to memorize the textbook or notes. Needing talent and good luck were moderately correlated ($r = .47$); these were summed and considered the level of *uncontrollable* attributions. Needing to study hard and to memorize notes were also correlated ($r = .37$); these were summed and considered the level of *controllable* attributions. Comparatively, the correlations between controllable and uncontrollable attributions were much smaller (.03 to .23).

There was no gender difference in self-efficacy or controllable attributions. However, the level of uncontrollable attributes was approximately 0.14 *SD*s higher for male students than for female students ($t = 5.7$, $p < .005$).

Student effort. Effort as reported by students was a complex characteristic. The amount of time students spent studying math after school on a normal school day was the only indicator (from several available in the TIMSS) with a reasonable amount of variance. About 17% reported to spend no time studying math whereas 57% spent less than 1 hr, 24% spent 1 to 2 hr, and 2% spent 3 or more hr.

In examining the relationship between three attitudinal items and time spent doing homework, another interpretation became evident (Figure 2). 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 1 hr or 1 to 2 hr on average. They also agreed to a greater extent that math was boring more than any other group. Students who studied more than 5 hr 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 level of skill as well as attitude toward mathematics, and was likely related to effort only among students who studied less than 3 hr on average. There was a nonlinear relationship between time spent on homework and achievement. This variable was transformed into two dummy variables to capture three

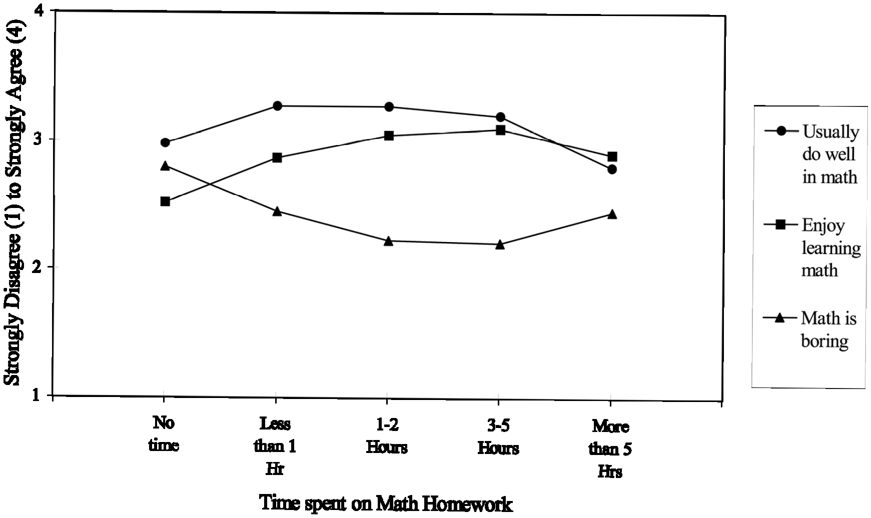

FIGURE 2 Attitudes toward mathematics by time spent on math homework.

levels of effort: (a) no homework, (b) more than 0 and up to 1 hr of homework, and (c) more than 1 hr of homework.

Additional student characteristics. Gender was coded 0 (*male*) or 1 (*female*). English was coded 0 (*sometimes or never speak English at home*) or 1 (*almost always or always speak English at home*). Mother's level of education was coded 1 (*finished primary school*), 2 (*finished some secondary school*), 3 (*finished secondary school*), 4 (*some vocational school*), 5 (*some university*), or 6 (*finished university*). Mother's expectation for mathematics performance was coded on a 4-point scale ranging from 1 (*never*) to 4 (*always*) as to whether students agreed with the following statement: My mother thinks it is important for me to do well in mathematics at school.

Measurement of Teacher Classroom Assessment Practices

Teacher classroom assessment practices were multifaceted and multidimensional. Teacher assessment practices were investigated for this project in two facets (Table 1). 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 including (a) the

TABLE 1
Classroom Assessment Practices

	Facets of Classroom Assessment	
Dimensions of Practice	*Homework Practices*	*Other Assessment Practices*
Tools	Workbook worksheets Textbook problems Textbook readings Writing assignments Data-collection activities Long-term individual projects Long-term small group projects Oral reports Journal writing	Teacher-made (T-M) objective (multiple-choice) tests T-M open-ended tests Projects Observations of students Student responses in class Externally created exams
Uses and related activities	Collect, correct, and keep Collect, correct, and return Contribute to grading Contribute to class discussion Give feedback to whole class Students correct their own work Students correct each others' work	Contribute to grading Feedback to students Give feedback to class Report to parents Group students Diagnose learning problems Plan future lessons

kinds of homework tasks that were assigned and (b) the uses teachers employ for the assigned homework.

All other assessment practices were included in the second facet of classroom assessment. 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—the uses teachers employed for the information obtained through their classroom assessment routines.

For each tool and use, teachers were asked to rate the frequency of assigning or employing each task in their classroom and the frequency in which they engaged in each use of assessment information on a 4-point scale ranging from XX (*never*) to XX (*always*).

Course content. A curriculum-related problem that TIMSS researchers have noted was the wide range of topics covered at all levels of rigor, thus, the characterization of a "splintered" curriculum, especially in the case of middle school mathematics courses (Schmidt, McKnight, & Raizen, 1996). There was no information available regarding the exact type of mathematics classes in which students were enrolled. However, teachers were asked to rate how much time they spent on each of 37 mathematics topics in their class during the year on a 5-point scale ranging from 0 (*not taught*) to 5 (*taught more than 15 periods during the year*).

Through a series of factor analyses, the 37 topics were combined into nine factors that closely reflected the logical combination of topics based on similarity of the form of mathematics employed. The nine factors accounted for 58% of the variance. Composite ratings for each factor were created by summing ratings for each topic within the resulting factor. Two factors seemed to differentiate the performance level of the classrooms, including time spent on (a) algebra (including the three topics of linear equations, inequalities, and formulas) and (b) fractions and whole numbers (eight topics, including common and decimal fractions and meaning of whole numbers, hereafter referred to as fractions). The time spent on algebra factor yielded the largest positive correlation with classroom achievement scores ($r = .44$), whereas the time spent on fractions factor yielded the largest negative correlation with achievement scores ($r = -.36$).

To provide a parsimonious indicator of relative prior math experience or content exposure of students within classrooms, the difference between time spent on algebra and time spent on fractions was referred to in this study as high–low relative prior math experience or simply *relative prior experience.* This variable 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). Relative prior experience of classrooms was normally distributed with a mean of 0.3 and standard deviation of 1.4 and was moderately related to classroom average mathematics scores ($r = .52$, $n = 328$).

HLM

HLM use has become an important tool in educational research where data are naturally nested, for example, when students are nested within classrooms (Frank, 1999). HLM appropriately accounts for the violation of the assumption of independence and two different degrees of freedom, one regarding the number of teachers and the second regarding the number of students. The HLM model was fit to the data based on the sample as described. This model was estimated using HLM 5.0 (Raudenbush, Bryk, & Congdon, 2000). Appropriate analyses were also conducted based on tests of coefficients, and model modifications were made. HLM also allows for the testing of significance in model-data fit between one or more models based on the inclusion or exclusion of certain estimated parameters. For a more complete description of estimation in HLM, see Bryk and Raudenbush (1992, pp. 32–56).

RESULTS

Teacher Classroom Assessment Practices

Homework. One facet of assessment practice included the two dimensions of homework: homework tasks and uses of homework. Few of the correlations between types and uses of homework were .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 ($r = .21$). Short writing tasks were assigned more often in classrooms in which teachers were more likely to collect and keep homework assignments ($r = .20$). Average classroom achievement was most highly correlated with the practice of correcting and keeping homework ($r = –.17$) and correcting and returning homework ($r = –.16$), all other correlations were less than .15. Use of textbook problems ($r = .23$), worksheets from workbooks ($r = –.18$), and reading tasks ($r = .13$) were weakly correlated with classroom achievement scores, whereas use of all other types of homework assignments were correlated with achievement less than .10.

Classroom assessment tools. The second facet of classroom assessment practice included the weight given to various tools by mathematics teachers to assess the work of their students. These various practices were intercorrelated at about .20 to .30, with the exception of externally created exams, which was weakly correlated with the others and observations of students, which was highly correlated with the use of student responses in class ($r = .79$). The use of teacher-made (T-M) objective tests was weakly, but negatively, correlated ($r = –.16$) with achievement scores, whereas all other tasks were correlated with achievement less than .10.

A second dimension in the classroom practice facet of assessment tools was the uses of those tools. These uses were moderately intercorrelated, between .21 and .55. The uses of assessment information were weakly correlated with achievement scores, all at .10 or less.

Most correlations between assessment tasks and uses of assessment information were small. The strongest relationships were among teachers who weight heavily T-M subjective tests and use assessment information for grading ($r = .20$), feedback ($r = .29$), and diagnosing learning problems ($r = .20$). Also, there were small relationships between teachers who give more weight to observations and responses of students and teachers who use assessment information for diagnosing problems and planning future lessons.

Relationships Between Teacher Practices and Classroom Level Achievement

Classroom level achievement. At the first stage, an unconditional HLM model (without explanatory variables) was specified and estimated, similar to a one-way analysis of variance with random effects, where classroom means were considered random and estimated from the mathematics scores of students nested within each classroom. This partitioned variance in achievement within and between classrooms. The unconditional model for mathematics achievement of students within classrooms was $Achievement_{ij} = \beta_{0j} + r_{ij}$ and $\beta_{0j} = \gamma_{00} + u_{0j}$, 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}) were 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}).

Based on the unconditional HLM model (Table 2), the maximum likelihood estimate of the grand mean (γ_{00}) mathematics achievement score was 0.017, essentially zero as expected due to the IRT scaling. The variance of classroom deviations (u_{0j}) from the grand mean was significantly different than zero, $\tau_{00} = 0.4835$, $\chi^2(320) = 8{,}781$, $p < .001$. Classrooms accounted for about 55% of the variance in student mathematics achievement performance (0.4835/[0.4835 + 0.3896]). 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% of the variance between schools (Bryk & Raudenbush, 1992). However, in this study, classrooms were 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 performance as compared to a classroom, particularly in a system in which students enroll in classes based on prior experience or exposure and skill or in which a high degree of tracking occurs.

Relative prior math experience. At the classroom level, the correlation between relative prior math experience (based on indicators described earlier) and the average achievement level of each class was evaluated earlier ($r = .52$). HLM analysis revealed the significance of the effect of relative prior math experience accounting for between-classroom variance.

Classrooms with higher relative prior math experience had higher average performance scores, as expected. This indicator provided an important control for prior experience based on prerequisite skill for the type of mathematics (content

TABLE 2
Unconditional Hierarchical Linear Modeling Model
of Student Mathematics Achievement Performance

Fixed Effects		*Coefficient*	*SE*	*T Ratio*	*p*
Intercept level 2, grand mean	γ^{00}	.017	.040	.438	.661
Random Effects		*Variance Component*	*df*	χ^2	*p*
Classroom mean residuals, u^{0j}	τ^{00}	.4835	320	8,781	.000
Student residuals, r^{ij}	σ_2	.3896			

and rigor) taught in a given classroom, and concomitantly the type of students enrolled in the class. Prior experience explained 28% of the between-classroom variance (after all other variables described below were added to the model, prior experience explained 13% of the between-classroom variance).

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, the student level mediating constructs were examined using a general linear model at the student level only. This step 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. After further evaluation, removing all nonsignificant interaction terms, only Mothers' Education × Uncontrollable Attributions was significant.

Three HLM models were assessed, each time removing nonsignificant terms and fixing level-1 slopes to be nonrandomly varying when the corresponding variance component was nonsignificant. The third and final model was

$$\textit{Achievement}_{ij} = \beta_{0j} + \beta_{1j}\,(\textit{Gender})_j + \beta_{2j}\,(\textit{English})_j + \beta_{3j}\,(\textit{Self-Efficacy})_j + \beta_{4j}\,(\textit{Uncontrollable Attributions})_j + \beta_{5j}\,(\textit{Uncontrollable Attributions} \times \textit{Mothers' Level of Education})_j + \beta_{6j}\,(<\textit{1-Hour Homework})_j + \beta_{7j}\,(\textit{No Homework})_j + \beta_{8j}\,(\textit{Mothers' Level of Education})_j + \beta_{9j}\,(\textit{Mothers' Expectations for Math Performance})_j + r_{ij}$$

$$\beta_{0j} = \gamma_{00} + \beta_{01}\,(\textit{Prior experience})_j + \gamma_{02}\,(\textit{Workbook Problems})_j + \gamma_{03}\,(\textit{T-M Objective Tests})_j + \gamma_{04}\,(\textit{Homework Frequency})_j + \gamma_{05}\,(\textit{Grade Level})_j + \gamma_{06}\,(\textit{Average Class Self-Efficacy})_j + \gamma_{07}\,(\textit{Average Class Uncontrollable Attributions})_j + \gamma_{08}\,(\textit{\% of Class that does No Homework})_j + u_{0j}$$

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

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

$$\beta_{3j} = \gamma_{30} + \gamma_{31}\,(\textit{T-M Objective Tests})_j + u_{3j}$$

$$\beta_{4j} = \gamma_{40} + \gamma_{41}\,(\textit{T-M Objective Tests})_j + \gamma_{42}\,(\textit{Prior experience})_j + u_{4j}$$

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

$$\beta_{6j} = \gamma_{60} + \gamma_{61}\,(\textit{Workbook Problems})_j$$

$$\beta_{7j} = \gamma_{70}$$

$$\beta_{8j} = \gamma_{80}$$

$$\beta_{9j} = \gamma_{90}$$

Estimates of each of the coefficients (γs) and random effects (τs) are presented in Table 3. The random effects are the residuals from each level 2 equation that

TABLE 3
Hierarchical Linear Modeling Model of Student Mathematics Achievement Given Classroom Level and Student Level Characteristics

Fixed Effects		*Coefficient*	*SE*	*T Ratio*	*p*
Model for classroom means, β_{0j}					
Intercept level 2, grand mean	γ_{00}	.016	.024	.659	.510
Prior experience	γ_{01}	.124	.020	6.185	.000
Workbook worksheets	γ_{02}	–.048	.038	–1.247	.213
T-M objective tests	γ_{03}	–.023	.031	–.750	.453
Homework frequency	γ_{04}	.082	.036	2.269	.023
Grade	γ_{05}	.151	.053	2.824	.005
Average self-efficacy	γ_{06}	.258	.097	2.654	.008
Average uncontrollable attribution	γ_{07}	–.811	.078	–10.431	.000
Percentage no homework	γ_{08}	–1.087	.209	–5.199	.000
Models for slopes					
Gender, β_{1j}	γ_{10}	–.106	.015	–7.077	.000
English, β_{2j}	γ_{20}	.064	.025	2.566	.011
Self-efficacy slope, β_{3j}	γ_{30}	.208	.012	16.884	.000
T-M objective tests	γ_{31}	–.035	.014	–2.439	.015
Uncontrollable atttributions slope, β_{4j}	γ_{40}	–.124	.019	–6.672	.000
T-M objective tests	γ_{41}	–.020	.011	–1.854	.063
Prior experience	γ_{42}	.012	.006	1.875	.060
Uncontrollable × Mothers' education slope, $\beta_{5\mathrm{j}}$	γ_{50}	.009	.004	2.058	.039
0–1 hr homework slope, β_{6j}	γ_{60}	.159	.018	9.026	.000
Workbook worksheets	γ_{61}	.077	.023	3.262	.001
No homework slope, β_{7j}	γ_{70}	.108	.025	4.361	.000
Mothers' education slope, β_{8j}	γ_{80}	.026	.005	4.802	.000
Mothers' expectations slope, β_{9j}	γ_{90}	.063	.014	4.505	.000

Random Effects		*Variance Component*	*df*	χ^2	*p*
Classroom mean residuals, u_{0j}	τ_{00}	.1730	312	3599	.000
Self-efficacy slope, u_{3j}	τ_{33}	.0060	319	363	.046
Uncontrollable attributions slope, u_{4j}	τ_{44}	.0048	318	386	.006
Student residuals, r_{ij}	σ^2	.3413			

vary significantly. In the case of u_{3j} and u_{4j}, the corresponding variances are the degree to which the slopes β_{3j} and β_{4j} vary across classrooms after adjusting for the explanatory variables in these equations.

The variables at the student level were centered on their classroom mean to retain the interpretation of the intercept: mean performance for the classroom, rather than an adjusted mean (the typical intercept interpretation). However, classroom mean centering 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. Those modeled only at the student level included mothers' education, mothers' expectations, and the interaction of uncontrollable attributions with mothers' education.

Overall, the model accounted for 64% of the variance in classroom means. Finally, the variance of the residual classroom means remained significant, $\chi^2(312) = 3{,}599$, $p < .001$.

All of the terms in the classroom level of the model were significant, except for the main effect of the use of T-M objective tests; however, the interactions between the use of T-M objective tests and student self-efficacy and uncontrollable attributions were significant.

Student characteristics. Gender had a significant but small effect (γ_{10} = –0.11), which suggested that female students scored slightly lower than male students, controlling for the other explanatory variables (i.e., all else constant). The magnitude of difference was about 0.11 of a standard deviation on the student mathematics score scale. This effect was constant across classrooms.

The effects of mothers' education level and mothers' expectations for performance in mathematics were statistically significant, positive, and constant across classrooms. The effect of having a mother who completed high school compared to those who completed college was smaller than the gender effect, 0.06 of a standard deviation. The effect of mothers' expectation to do well in math was also small (0.07 of a standard deviation).

The self-efficacy effect was dependent on the level with which teachers used T-M objective tests (γ_{31} = –0.035). Each level of use of objective tests from 0 (*none*) to 4 (*a great deal*) reduced the self-efficacy slope (γ_{30} = 0.208) by 0.035. Another way to interpret this cross-level interaction is that self-efficacy had a stronger relationship with math scores in classrooms where teachers did not heavily use T-M objective tests.

The effect of uncontrollable attributions by students (attributing success in mathematics to luck and natural talent) was dependent on the use of T-M objective tests (γ_{41} = –0.020) and relative prior experience level of the class (γ_{42} = 0.012). The use of T-M objective tests strengthened the negative relationship between uncontrollable attributions made by students and performance, whereas the higher prior experience level of the class weakened the negative relationship.

There was a statistically significant but very small interaction between the use of uncontrollable attributions and mothers' education level (γ_{50} = 0.009). It appeared that mothers' education level had a larger impact on math scores of students who rarely used uncontrollable attributions but less impact on scores of students who often used uncontrollable attributions.

Finally, the amount of time students spent on homework was related, in a restricted way, to math performance, all else constant. Students who did no homework each day performed slightly higher on average than those students who spent more than 1 hr a day on math homework ($\gamma_{70} = 0.108$). Again, as described earlier, the students who spent more than 1 hr each day studying mathematics were likely students who essentially needed to study more because of poor performance. Students who spent about 1 hr doing homework each day scored at an even higher level on average ($\gamma_{60} = 0.159$), about 0.17 *SD*s above the mean. The relationship between doing about 1 hr of homework and achievement was moderated by the use of workbook worksheets for homework ($\gamma_{61} = 0.077$). This result indicated that the relationship, or homework effect, was increased by 0.077 for each level of frequency of use of workbook problems, from 1 (*never*) to 4 (*always*); students in classrooms in which teachers assigned more workbook worksheets had stronger relationships between homework and achievement. These effects did not vary across classrooms and were unaffected by the level of homework assigned in general as reported by the teacher (tested in an earlier model).

Classroom characteristics and teacher assessment practices. The relative prior math experience indicator was a significant explanatory variable for classroom level performance ($\gamma_{01} = 0.124$), all else constant. Students enrolled in classes with the highest relative prior experience scored about 1.04 *SD*s above students in classes with the lowest relative prior experience. In addition, grade had a significant impact ($\gamma_{05} = 0.151$) as expected. Eighth-grade classrooms scored about 0.16 *SD*s above the average seventh-grade classroom, all else constant—including relative prior experience. Prior to conditioning on the other variables, students in eighth grade scored 0.36 *SD*s above students in seventh grade. This difference may be used as a guide-rule for comparing other differences, where we can consider 0.36 *SD*s difference on this TIMSS exam to be equivalent to 1 year of schooling.

The frequency with which teachers assigned homework (confounded with teachers who more frequently assigned text-book problem sets) had a significant relationship with math scores ($\gamma_{04} = 0.082$), all else constant. The classroom performance for those classrooms in which teachers assigned homework every day, compared to teachers who assigned homework once a week, was 0.18 *SD*s higher.

Frequent use of worksheets from workbooks as homework assignments had a negative relationship with classroom performance ($\gamma_{02} = -0.048$), all else constant. The difference in classroom performance for those classrooms in which the teacher never used workbooks versus teachers who always did so was 0.15 *SD*s. However, the use of T-M objective tests also had a small negative relationship with classroom performance ($\gamma_{03} = -0.023$), all else constant. The difference in classroom performance for those classrooms in which the teacher never used T-M objective tests versus teachers who did so frequently was 0.07 *SD*s.

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_{06} = 0.258), the average level of uncontrollable attributions made by students in a classroom (γ_{07} = –0.811), and the percent of students in the classroom who usually did no homework (γ_{08} = –1.087) were significant explanatory variables at the classroom level. After accounting for these differences among students within classrooms, their effect on between-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 0.47 *SD*s improvement in average class math performance due to overall classroom positive self-efficacy, 1.84 *SD*s improvement in average class math performance due to fewer overall classroom uncontrollable attributions, and 0.91 *SD*s drop in average class math performance due to students doing *no* homework.

Assessing the Adequacy of the HLM Model

The validity of inferences based on linear models in part depends on the defensibility of the assumptions of the model. In HLM, five key 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 (Bryk & Raudenbush, 1992).

One assumption regards the 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 across students within classrooms. Review of a normal probability plot of the level-1 residuals provided evidence of a fairly normal distribution, excluding slight deviation in the tails. However, the within-classroom variances were heterogeneous across classrooms, χ^2 (311) = 386, p = .003, violating the variance homogeneity assumption. The expected impact on the estimation of parameters and standard errors was minimal, partly because the value of the χ^2 was near the value of the degree of freedom. In addition, the restricted maximum likelihood pooled estimate of the variance as computed by HLM compensates for heterogeneity by increasing in size (Kasim & Raudenbush, 1998).

The remaining four assumptions regarding (a) independence of explanatory variables and error at level 1, (b) the distribution of error at level 2, (c) the independence of explanatory variables and error at level 2, and (d) the independence of errors between level 1 and level 2 were evaluated through methods recommended by Bryk and Raudenbush (1992). The assumptions appeared tenable and the model appeared appropriately specified. 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.

SUMMARY AND DISCUSSION

The primary research question used to direct this work is reviewed here explicitly: What are the interrelationships of teacher assessment practices, student characteristics, and achievement performance? Overall, the assessment practices of teachers were complex and not easily characterized. The use of homework tasks and various other assessment tools, and their purposes, were multifaceted. In short, classroom assessment practices were related to student performance and interacted in unique ways with student characteristics.

Given the unconditional HLM model (without explanatory variables), 55% of the variance in student performance scores was between-classrooms, whereas 45% was within-classrooms. The full HLM model with student and classroom level explanatory variables explained 64% of the variance between-classrooms and 8% of the variance within-classrooms.

The largest amount of variance explained was due to the prior experience indicator. As expected, the level of prior math experience of each class was a significant contributor to explaining variation in classroom performance and was viewed as an important control.

Several student level characteristics also differed on average across classrooms and contributed to the explanation of variation in classroom means. The average level of uncontrollable attributions (natural talent, luck) made by students in a classroom 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 large an impact as uncontrollable attributions. These are areas in which teachers have a potential to affect students: developing self-efficacy regarding potential for mastering mathematics and discouraging the uncontrollable attributions students make in the classroom (Brookhart, 1997; Glasser, 1985; Marsh & Craven, 1997; Wigfield, Eccles, & Rodriguez, 1998).

Classrooms in which large proportions of students did no homework were also classrooms in which teachers assigned less homework. However, even though classrooms in which teachers assigned frequent homework also had smaller proportions of students who did no homework, they also had students who reported spending no time to spending more than 3 hr a day on homework. Overall, both of these characteristics had significant independent effects; more frequent moderate levels of assigned homework were 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 in the form of workbook worksheets also improved the effect of doing about 1 hr of homework each day. That is, in classrooms where teachers assigned more worksheets, the positive effect of students doing homework about 1 hr on an average day was greater, all else equal—a result that complicates the overall negative effect of frequent use of workbook worksheets on classroom performance. Again, teachers may have some control over this practice. The range in percent of students in a classroom that did no homework was from 0% to 78% across classrooms. The resulting difference in classroom performance, all else constant, was more than 1 *SD* on the classroom average score scale, three times the difference between seventh- and eighth-grade performance.

These findings are in partial agreement with previous research. 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. Others have found significant relationships between homework and achievement (Cooper, 1989; Keith & Cool, 1992; Keith et al., 1993); however, Cooper suggested that this relationship might be curvilinear—as reported in these results.

Finally, certain classroom assessment practices were significantly related to classroom performance. Although the frequency of assigning homework had a positive relationship to performance, this frequency 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 they appear independent of the type of math class.

The use of T-M objective tests also had a negative relationship with mean classroom performance. As reviewed 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 T-M objective tests could result in lower performance on large-scale objective tests in a number of ways. Low quality tests do not provide valid indicators to students regarding their achievement and may inadvertently affect their academic self-efficacy, motivation, and effort.

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. Stiggins and Bridgeford (1985) and Stiggins and Conklin (1992) found higher use of T-M objective tests than were reported here. In all cases, T-M tests were more common than published or other written tests. Salmon-Cox (1980) also found that teachers relied on their own tests more than on interactions with students or homework. Regarding the use of assessment information, teachers use assessments largely to assign grades (Stiggins & Bridgeford, 1985; Stiggins & Conklin, 1992). Although eighth-grade teachers used T-M objective tests for diagnosis, grouping, evaluating, and reporting, math teachers relied more on T-M 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 the content and skills that they believe are important. If these things are poorly communicated, an expected 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. Students may learn as much from taking tests as any other activity in which they engage. This suggestion 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 instructional learning goals as well.

The impact of low-quality T-M 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 are low or high stakes.

To complicate matters even more, high reliance on T-M 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). For students in classrooms in which T-M objective tests were prevalent, the positive effect of self-efficacy was weaker than in classrooms in which T-M objective tests were not prevalent. Loosely speaking, greater focus on T-M 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 T-M tests as an assessment tool had indirect as well as direct negative relationships to student mathematics performance. A possibly confounding factor, as discussed earlier, could be quality of T-M objective tests.

These are areas in which 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: Use of T-M tests at the teacher level moderated the effect of self-efficacy and uncontrollable attributions at the student level; frequent worksheet homework tasks as assigned by teachers moderated the effect of students doing about 1 hr of homework a day. Unique combinations of teacher practices and student characteristics yield different results in terms of middle school mathematics performance.

Uncovering these unique interactions may help lead to more informed policy making regarding assessment reform efforts and the design of appropriate teacher training and professional development activities. 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 generalized model proposed here appeared tenable. Although the results of these correlational analyses do not provide evidence to support causal inferences, this study adds considerably to the level of complexity in considering the role of assessment practice and to defining some of the relationships involved in assessment practice.

REFERENCES

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

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

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5,* 7–74.

Blumenfeld, P. C., Puro, P., & Mergendoller, J. R. (1992). Classroom learning and motivation: Clarifying and expanding goal theory. *Journal of Educational Psychology, 84,* 272–281.

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

Brookhart, S. M., & DeVoge, J. G. (1999). Testing a theory about the role of classroom assessment in student motivation and achievement. *Applied Measurement in Education, 12,* 409–425.

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

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.

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

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

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

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

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

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

Gonzales, E. J., & Smith, T. A. (Eds.). (1997). *Users guide for the TIMSS International database*. Amsterdam, The Netherlands: IEA. Retrieved April 1, 1998, from http:// www.csteep.bc.edu/ timss1/database/ UG_1and2.pdf

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,* 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,* 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.

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.

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,* 321–342.

Nitko, A. J. (1989). Designing tests that are integrated with instruction. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 447–474). New York: American Council on Education & Macmillan.

Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2000). *HLM* (Version 5.0) [Computer software]. Chicago: Scientific Software International.

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

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

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25*, 173–180.

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

Stiggins, R. J. (1991). Facing the challenges of a new era of educational assessment. *Applied Measurement in Education, 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. (2001). *Student-involved classroom assessment* (3rd ed.). Upper Saddle River, NJ: Prentice-Hall, Inc.

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

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

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

Wigfield, A., Eccles, J. S., & Rodriguez, D. (1998). The development of children's motivation in school contexts. *Review of Research in Education, 23,* 73–118.