School Observation Measure

Reliability Study

Allan Sterbinsky

Steven M. Ross

July 2003

CREP
Center for Research in Educational Policy

School Observation Measure

Reliability Study

The recently enacted No Child Left Behind (NCLB) act emphasizes the use of scientifically based educational strategies and programs to significantly improve the academic achievement of students. This legislation is important because it funds reform efforts focused on improving the whole school as opposed to previous reform efforts that targeted individual teachers and subjects, which resulted in a patchwork of interventions with varying results (Ross et al. 2000). NCLB on the other hand, stimulates school-wide reform by covering virtually all aspects of school operations including instruction, assessment, classroom management, professional development, parental involvement, school management, and curriculum. A range of reform models that incorporate these elements are available for implementation at schools nationwide. Whole school reform models that are known to work in many locations usually incorporate a wide variety of classroom practices that are consonant with the philosophy of the model. Teachers typically receive training in the classroom practices as well as ongoing coaching to hone their skills while implementing these models.

Since teachers are the primary contact with students and have a high degree of control over the activities in which students engage, whole school reform models rightfully target classroom practices for change. It then follows that if classroom practices are the primary modality through which schools will improve, it is imperative that educators be able to reliably measure those classroom practices to determine if changes are indeed taking place across the whole school. In order to provide educators with the ability to measure classroom practices reliably and thus evaluate the ongoing implementation if the model (as required in NCLB) a reliable instrument capable of measuring a variety of classroom practices at the whole-school level must be used. Unfortunately, previous research in classroom observation measures has shown mixed results with regard to reliability.

Previous Research

Historically, measures of classroom practices have relied on classroom observations, but only at the teacher, student, or classroom level. In the 1970s' classroom observation instruments were used extensively in teacher effectiveness and teacher evaluation research, producing a body of literature on the reliability of classroom observation instruments. The logic of teacher effectiveness research assumed that specific classroom practices could be linked to increases in student outcomes. If these classroom practices are then used to evaluate teachers, teachers would be more likely to engage in those practices and student achievement would be more likely to increase (Rothenberg & Hessling, 1990).

A number of researchers during this time found significant linkages between classroom practices and student achievement (Karweit & Slavin, 1982; Brophy & Evertson, 1976; Flanders, 1970) while others found mixed results (Brophey & Good, 1986). Some researchers attributed the mixed results to the unreliability of classroom observation instruments in use at the time (Erlich & Shavelson 1978; Capie & Ellet 1982; Karweit & Slavin, 1982). They argued that if an instrument were unreliable, it is impossible for researchers to determine if there is no actually linkage between student achievement and classroom practices or if the potential linkages are obscured by the lack of instrument reliability. Researchers emphasized the need for conducting reliability studies before classroom observation instruments were used in research (Rowley, 1976; Rowley, 1978; Tobin & Capie, 1981; Karweit & Slavin, 1982). Unfortunately, systematic reliability studies of classroom observation instruments were, and still are rare (Marshall, Green, & Lawrence, 1976; Rothenberg & Hessling, 1990). An overview of the reliability studies available to date indicates that classroom observation instruments show evidence of reliability, but only in specific contexts.

Two well-known reviews of classroom observation instruments using observation instruments were conducted in the 1960's and 1970's and indicated a lack of reliability (Medley & Mitzel, 1963; Rosenshine, 1970). Other reliability studies (some conducted more recently) however, indicated more positive results. Lomax (1982) conducted a generalizability study with learning disabled students and found that the observation instrument yielded a stable measure of a range of classroom behaviors. Brophy, Coulter, Crawford, Evertson, and King (1975) found overall stability for many classroom variables but listed specific contexts in which the stability varied as did Marshall, Green, and Lawrence (1976). Rothenberg and Hessling (1990) reviewed reliability studies of observation instruments used specifically in North Carolina, Georgia, and Florida and found evidence for acceptable levels of reliability. In the majority of these studies, reliability estimates were tempered by the context of the observations.

These reviews indicate that contextual variables have a substantive impact on the reliability estimates provided for observation instruments. It is informative for our purposes to list the contextual variables that are most commonly mentioned in research of classroom observation instruments.

Contextual variables

Number and length of observations

- Cooley and Mau (1980) found that reliability estimates rose with the number of observations conducted.

- Tobin and Capie (1981) demonstrated that increasing either the number of observations conducted or the length of each observation increased the reliability "up to some asymptotic level."
- Rowley (1978) showed that increasing both the number of observations and the length of the observations increased the reliability estimates
- Karweit and Slavin (1982) noted that conventional wisdom indicated ten days of observations being enough to produce adequate reliability. The also noted that reducing the number of days of observation obscured the relationship between time-on-task, and reducing the number of students in the sample caused a small reduction in the reliability of the observational measures.

Frequency of occurrence of an item

- Marshall, Green, and Lawrence (1976) indicated that infrequently occurring items have low reliability estimates, but this is due to the lack of variability for infrequently occurring items. Some infrequently occurring items were actually very stable.
- Berliner (1976) stated that when the frequency of occurrence of an item is low, the reliability estimates for that item will also be low. He "found ratings of variables over 10 occasions that yield high stability coefficients" (p.8).

Type of items

- Dunkin and Biddle (1975) suggested combining low inference behavioral ratings into higher inference variables that reflect broader categories.

- Marshall, Green, and Lawrence (1976) stated that rating systems which relied on inferential judgments are "usually handicapped by the halo effect common to high-inference rating scales"(p2).
- Berliner (1976) opined that reliability estimates for ratings were higher than were those for frequency counts of behaviors
- Marshall (1975) states that "surface" aspects of the classroom such as materials or grouping of students may restrict rating scales, thus restricting variability.

Timing in academic year

- Berliner (1976) cited examples of classroom management behaviors that occur frequently during the first two weeks of the school year, but occur infrequently the remainder of the year. Observation schedules should take these differences into account.
- Evertson and Veldman (1981) studied changes in classroom behavior over time and found that midyear observations were less likely to be distorted than were those conducted during the early or late portions of the school year. The exceptions to this trend (yielding stable behaviors throughout the year) included time in seatwork behavioral, student initiated questions, and call-outs, which did not vary significantly throughout the course of the school year.

Changes in subjects or activities

- Marshall, Green, & Lawrence (1976) found that variability was associated with changes in classroom subjects or activities during the observation period.

Number of observers

- Padilla, Capie, and Cronin (1986) found that three observers were adequate to obtain sufficient reliability and the addition of another observer did not add substantively to the reliability estimate.

Type of observers

- Padilla, Capie, and Cronin (1986) found that external observers were more consistent in their scoring than were administrators. This indicates that administrators may not be as reliable as raters, but the actual differences in reliability between external raters and administrators were not great.

These variables have been shown to impact the reliability of classroom observations. As previous researchers have reiterated, it is essential to conduct reliability studies on the instruments before engaging in research. Otherwise, potential effects may be obscured by the lack of reliability inherent in the instrument and its use in schools.

Although these findings are important for the use of classroom observation instruments, one serious limitation remains as it relates to NCLB and the evaluations required by the legislation. Previous instruments focused on individual students, teachers, or classrooms. For purposes of NCLB, it is essential to measure changes at the whole-school level since the entire school is the target for change. Not only must the observation instrument be used at the school level, it must also measure a wide variety of classroom practices that may be implemented via the plethora of models adopted by schools. Currently, the only observation instrument that fulfills these requirements is the School Observation Measure.

The School Observation Measure (SOM)

The SOM was developed at the University of Memphis – Center for Research in Educational Policy for the purpose of measuring the extent to which different common and alternative classroom practices are used at the whole-school level. It measures the frequency of occurrence of 24 target practices and two summary items. Item categories include instructional orientation, classroom organization, instructional strategies, student activities, technology use, and assessment. The two summary items include academically focused class time and student attention/interest/focus. Observers participate in extensive training that includes observations in actual classrooms as well as consensus sessions that target interrater reliability.

The SOM measures classroom practices at the whole-school level, which makes it a very unique observation instrument. It was assumed that the contextual variables associated with student or teacher level observation instruments would hold true for the SOM as well. However, in agreement with the emphasis of Rowley, (1976); Rowley, (1978); Tobie and Capie, (1981); Karweit and Slavin, (1982), research must be conducted to determine the reliability of the SOM. For this reason, the current research study was conducted on data from approximately 1,100 SOMs conducted in 137 schools. Generalizability theory provided the theoretical foundations of the analysis and the familiar machinery of the ANOVA was used to estimate variance components due to different sources of variance in the data.

Participants

The current research study included approximately 1,100 SOMs conducted in 137 low performing schools in the southeast United States, which includes all schools in the United States that used the SOM during the 1999-2000 academic year. These schools typically received Federal funds to implement a whole-school reform model and served lower SES students (higher degree of poverty), with an average of 60% of all students being eligible for free or reduced price lunch. The number of students at each school ranged from 74 to 1,309, averaging approximately 500, and the student/ teacher ratio ranged from 9.4 to 31.6, with an average of 16 students per teacher. The sample contained 114 elementary schools (83%), 20 high schools (15%), 1 middle/high school (1%), 1 K-12 school (1%), and 1 junior high school (1%). Data from the 1999-2000 school year were used in this study because it was collected during the first year of whole-school reform implementation. The first year of whole-school reform typically consists of teacher training and planning sessions and relatively few reform-related changes actually being made in classroom practices. Whole-school reform related changes are usually more fully implemented during the second year of implementation, thus substantively changing classroom practices. For this reason, first year data provide a more stable baseline for measuring current classroom practices in schools. The number of SOMs conducted per school in this sample ranged from a low of two to a high of 12, with a mode of eight SOMs per school.

Measures

School Observation Measure (SOM)

The SOM was developed at the Center for Research in Educational Policy (CREP) at the University of Memphis and has been used in classrooms across the nation (Ross, et al. 1991). SOM observations consist of 15-minute visits to 10 randomly selected classrooms in a school. Observers record the frequency with which specific classroom practices are used during the 15-minute observation period. At the conclusion of the 10 observations, observers summarize the findings to produce an overall school-level rating for each SOM item. Summary findings rate the frequency with which each of the classroom practices were observed across all 10 classrooms. Items contained in the SOM include direct instruction, team teaching, cooperative/collaborative learning, individual tutoring, ability groups, multi-age grouping, work centers, higher level instructional feedback, integration of subject areas, project-based learning, higher-level questioning strategies, teacher acting as coach/facilitator, parent/community involvement, independent seatwork, experiential/hands-on learning, systematic individual instruction, sustained writing/composition, sustained reading, independent inquiry/research, student discussion, computer for instructional technology, technology as learning tool/resource, performance assessment strategies, student self-assessment, academically focused class time, and student attention/interest/focus. Possible ratings for each category include not observed, rarely, occasionally, frequently, or extensively observed (see Appendix A). The two summary items (academic focus and student attention) are rated on a scale of 1 (low) to 3 (high).

The use of 10 classroom observations as the basis for one SOM effectively constitutes 150 minutes of observation for each SOM conducted. Using eight SOMs for each school equates to 1,200 minutes of observation time per school. This approach to collecting observation data provides extensive observation time in classes school-wide, thus addressing the concerns raised by Cooley and Mau (1980), Tubin and Capie (1981), Rowley (1978), and Karweit and Slavin (1982) regarding the number and length of observations.

Overall ratings on the SOM (after observing 10 classes) are based on low-inference notations completed in each of the classrooms. Classroom notes are used in toto for purposes of rating the frequency of occurrence of each item. This procedure addresses the issue of high versus low inference items raised by Duncan & Biddle (1975), Everson and Veldman (1976), Marshall, Green, and Lawrence (1976), and Berliner (1976). Additionally, some of the SOM items measure what Marshall (1975) calls “surface” aspects of the classroom (e.g. grouping of students).

Procedures

During the summer of 1999, SOM observers (external observers) were trained in the use of the SOM by the SOM developers. These observers were typically retired educators with multiple years’ experience. The use of external observers was based in part on the findings of Padilla, et al. (1986) regarding the higher reliability coefficients generated by external observers versus school administrators.

During the 1999-2000 school year, observers visited each school multiple times to conduct SOM observations in randomly selected classrooms. SOMs were typically conducted in the fall (October) and continued periodically throughout the school year, with the last SOM being conducted before the last month of the school year. This observation schedule ameliorates the concerns raised by Berliner, (1976) and Evertson and Veldman (1981) regarding differences in behaviors during the first and last months of the academic year.

Multiple observers completed SOMs at the schools included in this study, thus addressing the concerns of Padilla, Capie, and Cronin (1986) related to the differences between observers. Due to the wide geographic dispersion of schools, no attempt was made to fully cross raters with schools. One rater may have completed all SOMs at some schools, while at other schools, multiple raters may have completed the SOMs. Since this is reflective of the reality of SOM use under normal conditions, the reliability study was conducted under similar conditions to provide an accurate assessment of reliability under normal conditions.

Since contextual variables impact reliability coefficients, it was assumed that school demographics may also impact the coefficients in this study. For this reason, school demographics are reported for the schools in the sample. School demographic data were obtained from the National Center for Education Statistics (NCES), which is the primary federal entity for educational statistics under the auspices of the U.S. Department of Education. NCES provides data such as the total number of students at each school (school size) and the student-teacher ratio, which is a computation of the total number of full-time equivalent teachers divided by the total number of students at the school.

Degree of poverty was defined by NCES as the percentage of students at each school who were eligible for free or reduced price lunch under the National School Lunch Act. NCES also identified the locale of each school, ranging from a large central city (population greater than or equal to 250,000) to rural (population less than 2,500). Schools from rural or small town locations were categorized as rural schools. Schools from large town or urban areas were categorized as urban schools.

The reliability of the SOM was estimated using SOM data from all schools at which at least eight SOMs were conducted (*n* = 116). This criteria ensures that sufficient SOM data are available at each school to provide an accurate estimate of the reliability of the SOM. For schools having more than eight SOMs conducted during the school year, eight SOMs were randomly selected from the total number of SOMs conducted at the school for analysis purposes. Norms for each classroom practice are reported in Appendix B and were based on the full sample of schools in the study (N=137).

Data Analysis

Traditionally, classical test theory has been used to estimate the reliability of instruments and continues to be used extensively (Suen, 1990). To estimate the test-retest reliability of the SOM using classical methods, a reliability coefficient would be calculated using two separate occasions of SOM observations across a number of schools. This coefficient would be an estimate the degree of relationship between the first and second SOM observations. This approach is meaningful when there are only two occasions of SOM observations. In reality however, multiple SOMs are typically conducted at schools, consisting of anywhere between five to ten SOMs. To estimate the test-retest reliability of multiple pairs of SOMs using traditional methods, one pair of SOMs out of all possible SOM pairs would be selected for calculating the reliability of the instrument. The difficulty with this approach is that reliability estimates from the same dataset will yield different estimates depending on which pair of SOMs was chosen for analysis. An answer to this concern would be to calculate reliability coefficients for every possible pair of SOMs and report the average as the final reliability coefficient. Although this approach has advantages over the first approach, it inherently contains three limitations. First, the average reliability coefficient does not address all SOM observations simultaneously. Secondly, there is only one undifferentiated error term in the traditional reliability coefficient formula, which does not allow researchers to address the many potential sources of error that exist in observational measures. Thirdly, use of the traditional reliability coefficient is only capable of estimating relative rankings of subjects and not absolute standards.

Recognizing these limitations in the traditional approach to test-retest reliability, Cronbach, Gleser, Nanda, and Rajaratnam (1972) approached reliability via a different conceptual framework called Generalizability Theory. They assumed that “an investigator asks about the precision or reliability of a measure because he wishes to generalize from the observation in hand to some class of observations to which it belongs” (p. 15). Thus, the classical definition of reliability is replaced by a broader question, which asks how accurately observed scores permit generalization of behavior in a defined universe (Shavelson & Webb, 1991; Shavelson, Webb, & Rowley, 1989).

Generalizability theory (GT) is composed of two phases: 1) a G (Generalizability) study, which estimates variability due to SOMs based on current SOM data and, 2) a D (Decision) study which applies the results of the G study to a range of possible future scenarios.

G study

A G study (as used in the current study) collects SOM data on multiple occasions at each school. It is assumed that these schools are reasonably representative of all schools that will use the SOM at some point in the future. SOM data collected from these schools are then analyzed using an analysis of variance (ANOVA) for the purpose of providing variance estimates from each source of variance. The independent variables in the ANOVA for the current study include schools and SOM occasions (each SOM conducted at a school represents one occasion). The ANOVA decomposes observed SOM scores into variance components due to different sources of variability including schools (*n* = 116), occasions (eight SOMs conducted at each school), and the interaction term (School by SOM). Mean squares from each source of variance are then used to calculate estimated variance components for SOMs using the formulas in Table 1. In the formulas, the subscript *p* indicates the variance associated with schools, the subscript *i* indicates the variance associated with occasions of SOM observations, and the subscript *pi, e* indicates the error term which is composed of the interaction of occasions and schools. This approach allows all SOMs to be addressed simultaneously as well as estimating multiple sources of variance (occasions, or others), and estimations using absolute standards.

In this study, the universe of observers consists of all people who successfully complete the SOM training. For this reason, generalizations can be made over all such people. The universe of schools is random since all schools in the US were not included in the study.

Table 1

Sources of Variance and Formulas For Estimating Variance Components

Source of Variance	Sum of Squares	*Df*	Mean Square	Estimated Variance
School ($\hat{\sigma}_p^2$)				$\frac{MS_p - \hat{\sigma}_{pi,e}^2}{n_i}$
SOM Occasions ($\hat{\sigma}_i^2$)				$\frac{MS_i - \hat{\sigma}_{pi,e}^2}{n_p}$
Occasions by School ($\hat{\sigma}_{pi,e}^2$) (Error term)				$MS_{pi,e}$

Calculations of the estimated variance proceed from the bottom of the traditional ANOVA table to the top. Since the model is fully defined by schools and SOM observations, there is only one observation per cell. For this reason, the error term consists of the SOM by School interaction term. The estimated variance component for the error term is simply the mean square of the interaction term. The estimated error variance associated with the SOM is defined as the mean square of the SOM term, from which the error term is subtracted. This value is then divided by the number of schools in the equation. A similar procedure is used to calculate the estimated variance for the school term. The variance estimates provided by the ANOVA in the G study are then used in the D study, which allows researchers to estimate the reliability of the SOM under specific future conditions.

Another helpful statistic provided by the G study is percent of variance. Each source of variance in the ANOVA yields a variability estimate attributable to that source of variance. Variances from each source are combined to yield the total variance. The total variance estimate from the ANOVA is then used to determine the relative amount of variance associated with each source of variance by dividing the variance estimate from each source of variance by the total variance. The resultant percent of variance statistic helps to interpret the relative amount of variance associated with each source of variance. For example, if the percentage of variance attributed by schools was 80%, the variance due to SOMs was 5%, and the variance due to the interaction (error) term was 15%, this would indicate that schools were substantively different from each other, the statistical model was fairly well defined (relatively low percentage of variance), and that the SOMs did not vary much from each other. If however, the percentage of variance attributable to the error term was 80%, this would indicate that the statistical model was not well defined and that other sources of variance should be included in the model in order to control for those sources of error.

D study

Once variance estimates are completed via the G study using actual data from schools, they can then be applied to hypothetical situations (e.g., future studies using the SOM). The application of variance estimates from the G study to hypothetical situations is called a D study. In the current example, the D study used actual data from 8 SOMs collected from 137 schools to produce variance estimates. These variance estimates are then applied to hypothetical situations wherein schools could use only 3, 5, 6, 10, or 20 SOMs. The phi coefficient generated by the D study is analogous to the reliability coefficient in traditional reliability studies, and a separate phi coefficient is generated for each hypothetical number of SOMs that could be used in a school. For example, a phi coefficient hypothetically based on five SOMs will use the variance estimates based on eight SOMs to generate a phi coefficient associated with potentially using only five SOMs.

The first step in a D study is to determine the conditions under which the SOM will be used in future scenarios. These decisions include an absolute versus relative standard, and the number of SOMs that could be conducted at schools in future scenarios. Once these decisions are made, a phi coefficient (analogous to a reliability coefficient) is calculated for each future scenario.

For purposes of the current study, it was assumed that the SOM would be used to make absolute decisions as opposed to relative ones. A relative decision would compare schools based on their use of SOM items relative to other schools. Since the SOM is used primarily to determine the absolute frequency with which specific classroom practices occur within a school (regardless of their relative ranking compared to other schools), an absolute standard was used in the D study. The variance formula associated with absolute decisions is provided by Shavelson and Webb (1991). In this formula, the subscript *p* indicates the variance associated with schools, the subscript *i* indicates the variance associated with occasions of SOM observations, and the subscript *pi*, *e* indicates the error term which is composed of the interaction of occasions and schools.

$$\hat{\sigma}^2_{Abs} = \frac{\hat{\sigma}^2_i}{n_i} + \frac{\hat{\sigma}^2_{pi,e}}{n_i}$$

The next step in a D study is to determine the number of possible SOMs that could be conducted at a school and calculate phi coefficients (analogous to a reliability coefficient) based on the number of SOMs that could be used. For example, if a school chooses to conduct only one SOM, a phi coefficient can be calculated for the use of one SOM (based on the variance estimates provided in the G study), which will estimate the degree to which that one SOM will generalize to the universe of all possible SOMs that could have been conducted at the school. If a school chooses to conduct 20 SOMs, another phi coefficient can be calculated (based the assumption of 20 SOMs), which will estimate the degree to which the mean of those 20 SOMs will generalize to the universe of all possible SOMs that could have been conducted at that school.

The formula for the phi coefficient (Kane & Brennan, 1977) is:

$$\phi = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_{Abs}^2}$$

The phi coefficient is analogous to a reliability coefficient, and if only one source of variance is included in the study. If a relative decision is made, the resultant phi coefficient is the same as the coefficient produced by classic test-retest reliability. For the current study, it was assumed that future scenarios would include schools using 3, 5, 6, 8, 10, and 20 SOMs per academic year.

Results

GT is composed of a G study and a D study. Results from both the G study and D study in the current research are presented in separate sections.

G study

An ANOVA was used to decompose SOM scores into variance components, including schools, SOMs, and the error term (interaction). Mean squares from the ANOVA were then used to estimate the variance components using the formula described in the methods section. As seen in Table 2, the variance estimates attributable to the SOMs were substantively smaller than the variance estimates attributable to schools or the error term, ranging from a low of less than .0000 to a high of .0042. The percentage of variance attributable to each source of variance was also calculated by dividing the variance attributable to each source of variance by the total variance. The percent of variance attributable to SOM occurrences ranged from less than .000% to 1.42% (*M* = .17%) while the variance attributable to schools ranged from 23 to 64% (*M* = 38%). This indicates that differences between SOM occasions were relatively minor, and differences between schools were substantively higher. The overall percentage of variance attributable to the error term (SOM occasions by school) ranged from 36 to 75% (*M* = 62%), much higher than for either SOM occasions or schools. This indicates that a substantive percent of error variance is unexplained by the model. Potential sources of variance could include raters, time of day, and day of the week.

Table 2

Variance Estimates and Percentage of Variance Attributable To SOMs, Schools, and Error Terms (n = 116 Schools)

	Variance Estimates			Percentage of Variance		
	School	SOM	Error	School	SOM	Error
Instructional Orientation						
Direct Instruction	0.1833	0.0033	0.5970	23.39	0.43	76.19
Team Teaching	0.2999	0.0004	0.3310	47.50	0.06	52.43
Cooperative/collaborative learning	0.3380	0.0020	0.5770	37.01	0.00	63.18
Individual tutoring	0.3428	0.0000	0.5560	38.18	0.00	61.93
Classroom Organization						
Ability groups	1.0398	0.0037	0.7370	58.40	0.21	41.39
Multiage grouping	0.9311	0.0028	0.5180	64.13	0.19	35.68
Workcenters	0.4155	0.0001	0.6060	40.73	0.00	59.40
Instructional Strategies						
Higher-level instructional feedback	0.5641	0.0018	0.9400	37.46	0.12	62.42
Integration of subject areas	0.2428	0.0000	0.4050	37.53	0.00	62.61
Project-based learning	0.1230	0.0000	0.3060	28.79	0.00	71.62
Higher-level questioning	0.3738	0.0010	0.6090	38.08	0.00	62.05
Teacher as coach	0.5193	0.0000	0.8130	39.00	0.00	61.06
Parent/community involvement	0.0939	0.0042	0.1980	31.70	1.42	66.88
Student Activities						
Independent seatwork	0.2264	0.0019	0.6580	25.54	0.21	74.24
Experiential, hands-on learning	0.3003	0.0000	0.5240	36.47	0.00	63.65
Systematic individual instruction	0.1886	0.0000	0.2660	41.52	0.00	58.55
Sustained writing	0.1363	0.0022	0.3500	27.90	0.44	71.66
Sustained reading	0.2631	0.0001	0.5350	33.02	0.00	67.14
Independent inquiry	0.0823	0.0006	0.2190	27.25	0.19	72.56
Student discussion	0.3448	0.0000	0.5360	39.29	0.37	61.08
Technology Use						
Computer for instructional delivery	0.2604	0.0037	0.4180	38.18	0.54	61.29
Technology as a learning tool	0.2035	0.0000	0.3610	36.09	0.00	64.02
Assessment						
Performance assessment strategies	0.2484	0.0014	0.4370	36.17	0.20	63.63
Student self-assessment	0.1890	0.0000	0.2590	42.31	0.00	57.98
Summary Items						
Academically focused class time	0.1031	0.0000	0.2010	34.01	0.00	66.29
Student attention/interest/focus	0.1235	0.0000	0.2000	38.18	0.00	61.83
Mean	0.3130	0.0011	0.4676	37.61	0.17	62.34

D study

The next step in Generalizability theory is to conduct a D study, which uses the results of the G study and applies them to potential future scenarios. For example, a D study uses the coefficients generated in the G study (based on schools that had eight SOMs) to estimate what the reliability would be if schools used a different number of SOMs (e.g., 1, 3, 5, 6, 8, 10, or 20). The coefficients for eight SOMs are based on actual data generated in the G study, however the reliability estimates for any different number of SOMs at a school are based on extrapolations using the coefficients generated in the D study.

For the purposes of this study, it was assumed that schools would choose to use 1, 3, 5, 6, 8, 10, or 20 SOMs, and a separate phi coefficient was calculated based on each of these scenarios. Table 3 lists the ϕ coefficient for each SOM item under each of the assumed future scenarios. The ϕ coefficients averaged across all SOM items varied by the number of SOMs conducted at a school. Conducting just one SOM observation in a school yielded a low generalizability measure (ϕ = .38), indicating that the results of one SOM conducted at the school would not sufficiently generalize to the set of all possible SOMs that could have been conducted at the school. Increasing the number of SOMs at a school to three SOMs increased the generalizability on average to .64, and increasing the number of SOMs to eight at a school increased the generalizability on average to .82, a more reasonable level of estimated generalizability.

It should be noted that three items evidence relatively high levels of reliability even when only one SOM is conducted (team teaching = .48, ability groups = .58, and multiage grouping = .64). These results are probably due to the fact that SOM observers must ask follow-up questions regarding these three items. For example, if an observer is in a primary school, it is difficult to tell simply by observation if some children in the room are second graders (smaller than the other children) or third graders (bigger than the other children), which necessitates observers directly asking teachers if multiage grouping is being used. The same is true for ability grouping. Since observers are unable to determine accurately if a small group of students is composed of similar or disparate ability levels, the observers must ask the teachers. Finally, if two adults are in the classroom, observers must ask the teachers if both adults are teachers, or if one adult is a volunteer or an aide.

Phi coefficients were also calculated separately for rural and urban elementary schools, and high schools to determine if reliability estimates differ by these categories. On average, conducting five SOMs at urban elementary schools yielded a ϕ coefficient of .72 (as seen in Table 4), which was similar to the coefficient based on all schools ($\phi = .74$). Similarly, conducting five SOMs at a rural elementary school yielded a ϕ coefficient of .76 (as seen in Table 5). Conducting five SOMs at a high school, however, yielded a ϕ coefficient of .60, which is somewhat lower than that obtained at elementary schools (see Table 6). For high schools, increasing the number of SOMs to eight would yield the same level of generalizability as five would in the elementary schools. Since the ϕ coefficients for high schools were based on ANOVA data from only 20 high schools, the low ϕ coefficient may be due to the low number of schools used in the G study subsample rather than any inherent unreliability of SOMs conducted at high schools. These results indicate that a minimum of five SOMs would provide a minimal level of generalizability to the set of all SOMs that could have been conducted at a school.

Table 3

Phi coefficients by SOM Item associated with number of SOM occasions in the D study (All Schools: n = 116)

	Number of SOM Occasions in D study						
SOM Item	1	3	5	6	8	10	20
Instructional Orientation							
Direct Instruction	0.23	0.48	0.60	0.65	0.71	0.75	0.86
Team Teaching	0.48	0.73	0.82	0.84	0.88	0.90	0.95
Cooperative/collaborative learning	0.37	0.64	0.75	0.78	0.82	0.85	0.92
Individual tutoring	0.38	0.65	0.76	0.79	0.83	0.86	0.93
Classroom Organization							
Ability groups	0.58	0.81	0.88	0.89	0.92	0.93	0.97
Multiage grouping	0.64	0.84	0.90	0.91	0.93	0.95	0.97
Workcenters	0.41	0.67	0.77	0.80	0.85	0.87	0.93
Instructional Strategies							
Higher-level instructional feedback	0.37	0.64	0.75	0.78	0.83	0.86	0.92
Integration of subject areas	0.38	0.64	0.75	0.78	0.83	0.86	0.92
Project-based learning	0.29	0.55	0.67	0.71	0.76	0.80	0.89
Higher-level questioning	0.38	0.65	0.75	0.79	0.83	0.86	0.92
Teacher as coach	0.39	0.66	0.76	0.79	0.84	0.86	0.93
Parent/community involvement	0.32	0.58	0.70	0.74	0.79	0.82	0.90
Student Activities							
Independent seatwork	0.26	0.51	0.63	0.67	0.73	0.77	0.87
Experiential, hands-on learning	0.36	0.63	0.74	0.77	0.82	0.85	0.92
Systematic individual instruction	0.42	0.68	0.78	0.81	0.85	0.88	0.93
Sustained writing	0.28	0.54	0.66	0.70	0.76	0.79	0.89
Sustained reading	0.33	0.60	0.71	0.75	0.80	0.83	0.91
Independent inquiry	0.27	0.53	0.65	0.69	0.75	0.79	0.88
Student discussion	0.39	0.66	0.76	0.80	0.84	0.87	0.93
Technology Use							
Computer for instructional delivery	0.38	0.65	0.76	0.79	0.83	0.86	0.93
Technology as a learning tool	0.36	0.63	0.74	0.77	0.82	0.85	0.92
Assessment							
Performance assessment strategies	0.36	0.63	0.74	0.77	0.82	0.85	0.92
Student self-assessment	0.42	0.69	0.79	0.81	0.85	0.88	0.94
Summary Items							
Academically focused class time	0.34	0.61	0.72	0.76	0.81	0.84	0.91
Student attention/interest/focus	0.38	0.65	0.76	0.79	0.83	0.86	0.93
Mean across all SOM items	0.38	0.64	0.74	0.77	0.82	0.85	0.92

Table 4

Phi coefficients by SOM Item associated with number of SOM occasions in the D study (Urban Elementary Schools: n = 34)

	Number of SOM Occasions in D study						
SOM Item	1	3	5	6	8	10	20
Instructional Orientation							
Direct Instruction	0.16	0.36	0.48	0.53	0.60	0.65	0.79
Team Teaching	0.47	0.73	0.82	0.84	0.88	0.90	0.95
Cooperative/collaborative learning	0.47	0.73	0.82	0.84	0.88	0.90	0.95
Individual tutoring	0.42	0.69	0.78	0.81	0.85	0.88	0.94
Classroom Organization							
Ability groups	0.56	0.79	0.86	0.88	0.91	0.93	0.96
Multiage grouping	0.54	0.78	0.85	0.87	0.90	0.92	0.96
Workcenters	0.24	0.49	0.61	0.65	0.72	0.76	0.86
Instructional Strategies							
Higher-level instructional feedback	0.59	0.81	0.88	0.90	0.92	0.94	0.97
Integration of subject areas	0.41	0.67	0.78	0.81	0.85	0.87	0.93
Project-based learning	0.28	0.54	0.67	0.70	0.76	0.80	0.89
Higher-level questioning	0.44	0.70	0.80	0.82	0.86	0.89	0.94
Teacher as coach	0.46	0.72	0.81	0.84	0.87	0.89	0.94
Parent/community involvement	0.13	0.30	0.42	0.46	0.54	0.59	0.74
Student Activities							
Independent seatwork	0.35	0.62	0.73	0.76	0.81	0.84	0.92
Experiential, hands-on learning	0.32	0.59	0.71	0.74	0.79	0.83	0.91
Systematic individual instruction	0.40	0.67	0.77	0.80	0.84	0.87	0.93
Sustained writing	0.28	0.53	0.66	0.70	0.75	0.79	0.88
Sustained reading	0.28	0.53	0.65	0.69	0.75	0.79	0.88
Independent inquiry	0.13	0.31	0.43	0.48	0.55	0.61	0.75
Student discussion	0.47	0.73	0.82	0.84	0.88	0.90	0.95
Technology Use							
Computer for instructional delivery	0.45	0.71	0.81	0.83	0.87	0.89	0.94
Technology as a learning tool	0.35	0.62	0.73	0.76	0.81	0.84	0.91
Assessment							
Performance assessment strategies	0.43	0.69	0.79	0.82	0.86	0.88	0.94
Student self-assessment	0.40	0.67	0.77	0.80	0.84	0.87	0.93
Summary Items							
Academically focused class time	0.23	0.47	0.60	0.64	0.70	0.75	0.86
Student attention/interest/focus	0.27	0.52	0.65	0.69	0.75	0.79	0.88
Mean across all SOM items	0.37	0.61	0.72	0.75	0.80	0.83	0.90

Table 5

Phi coefficients by SOM Item associated with number of SOM occasions in the D study (Rural Elementary Schools: n = 58)

	Number of SOM Occasions in D study						
SOM Item	1	3	5	6	8	10	20
Instructional Orientation							
Direct Instruction	0.30	0.56	0.68	0.72	0.77	0.81	0.90
Team Teaching	0.39	0.66	0.76	0.80	0.84	0.87	0.93
Cooperative/collaborative learning	0.38	0.65	0.75	0.79	0.83	0.86	0.92
Individual tutoring	0.35	0.62	0.73	0.77	0.81	0.85	0.92
Classroom Organization							
Ability groups	0.59	0.81	0.88	0.90	0.92	0.93	0.97
Multiage grouping	0.71	0.88	0.93	0.94	0.95	0.96	0.98
Workcenters	0.40	0.66	0.77	0.80	0.84	0.87	0.93
Instructional Strategies							
Higher-level instructional feedback	0.34	0.61	0.72	0.76	0.81	0.84	0.91
Integration of subject areas	0.38	0.65	0.75	0.79	0.83	0.86	0.92
Project-based learning	0.34	0.61	0.72	0.76	0.80	0.84	0.91
Higher-level questioning	0.40	0.67	0.77	0.80	0.84	0.87	0.93
Teacher as coach	0.43	0.69	0.79	0.82	0.86	0.88	0.94
Parent/community involvement	0.41	0.67	0.77	0.80	0.85	0.87	0.93
Student Activities							
Independent seatwork	0.24	0.49	0.61	0.66	0.72	0.76	0.86
Experiential, hands-on learning	0.41	0.67	0.77	0.80	0.85	0.87	0.93
Systematic individual instruction	0.45	0.71	0.80	0.83	0.87	0.89	0.94
Sustained writing	0.30	0.56	0.68	0.72	0.77	0.81	0.89
Sustained reading	0.29	0.55	0.67	0.71	0.76	0.80	0.89
Independent inquiry	0.40	0.66	0.77	0.80	0.84	0.87	0.93
Student discussion	0.38	0.65	0.75	0.79	0.83	0.86	0.92
Technology Use							
Computer for instructional delivery	0.31	0.58	0.69	0.73	0.78	0.82	0.90
Technology as a learning tool	0.37	0.64	0.74	0.78	0.82	0.85	0.92
Assessment							
Performance assessment strategies	0.34	0.61	0.72	0.76	0.80	0.84	0.91
Student self-assessment	0.49	0.74	0.83	0.85	0.88	0.91	0.95
Summary Items							
Academically focused class time	0.37	0.64	0.74	0.78	0.82	0.85	0.92
Student attention/interest/focus	0.42	0.69	0.79	0.81	0.85	0.88	0.94
Mean across all SOM items	0.40	0.65	0.76	0.79	0.83	0.86	0.92

Table 6

Phi coefficients by SOM Item associated with number of SOM occasions in the D study (All High Schools: n = 20)

	Number of SOM Occasions in D study						
SOM Item	1	3	5	6	8	10	20
Instructional Orientation							
Direct Instruction	0.05	0.13	0.19	0.22	0.28	0.32	0.49
Team Teaching	0.37	0.64	0.75	0.78	0.83	0.86	0.92
Cooperative/collaborative learning	0.11	0.28	0.39	0.43	0.50	0.56	0.72
Individual tutoring	0.46	0.72	0.81	0.84	0.87	0.89	0.94
Classroom Organization							
Ability groups	0.67	0.86	0.91	0.92	0.94	0.95	0.98
Multiage grouping	0.56	0.79	0.86	0.88	0.91	0.93	0.96
Workcenters	0.36	0.63	0.74	0.77	0.82	0.85	0.92
Instructional Strategies							
Higher-level instructional feedback	0.17	0.39	0.51	0.56	0.63	0.68	0.81
Integration of subject areas	0.24	0.49	0.61	0.65	0.72	0.76	0.86
Project-based learning	0.17	0.37	0.50	0.54	0.61	0.66	0.80
Higher-level questioning	0.17	0.38	0.50	0.55	0.62	0.67	0.80
Teacher as coach	0.22	0.45	0.58	0.63	0.69	0.74	0.85
Parent/community involvement	0.05	0.13	0.20	0.23	0.28	0.33	0.50
Student Activities							
Independent seatwork	0.15	0.35	0.47	0.51	0.59	0.64	0.78
Experiential, hands-on learning	0.37	0.64	0.74	0.78	0.82	0.85	0.92
Systematic individual instruction	0.28	0.54	0.66	0.70	0.76	0.80	0.89
Sustained writing	0.22	0.46	0.59	0.63	0.69	0.74	0.85
Sustained reading	0.24	0.49	0.61	0.66	0.72	0.76	0.86
Independent inquiry	0.16	0.36	0.49	0.53	0.60	0.65	0.79
Student discussion	0.34	0.61	0.72	0.76	0.80	0.84	0.91
Technology Use							
Computer for instructional delivery	0.16	0.37	0.49	0.54	0.61	0.66	0.80
Technology as a learning tool	0.23	0.47	0.60	0.64	0.71	0.75	0.86
Assessment							
Performance assessment strategies	0.29	0.55	0.67	0.71	0.76	0.80	0.89
Student self-assessment	0.24	0.48	0.61	0.65	0.71	0.76	0.86
Summary Items							
Academically focused class time	0.41	0.68	0.78	0.81	0.85	0.87	0.93
Student attention/interest/focus	0.36	0.62	0.74	0.77	0.82	0.85	0.92
Mean across all SOM items	0.27	0.50	0.60	0.64	0.70	0.74	0.84

Summary and Conclusion

The recently enacted NCLB legislation funds the implementation of whole-school reforms and requires evaluation of these reforms by educators. Since whole-school reforms typically change practices in the classroom, educators need a reliable classroom observation instrument that measures a wide variety of classroom practices. Historically, classroom observation instruments were designed for individual students, teachers, or classrooms, not the whole school. Additionally, the reliability of those instruments was shown to be context specific (e.g. length of observation, number of observations, timing in academic year).

The SOM is designed to measure a variety of classroom practices at the school level and the current reliability study was designed to incorporate many of the contextual variables elucidated by previous research. The current research study provides ample evidence of reliability for the SOM in a range of contexts. On average, the phi coefficient across all SOM items was .74 for five observations and .82 for eight observations at a school. Additionally, the percentage of variance attributable to observations was less than 1% while the percentage of variance attributable to differences between schools was 37% and the error term was 62%.

Disaggregating reliability coefficients by elementary/high schools and rural/urban locations also evidenced reliability across contexts. The phi coefficients for five SOMs conducted at urban elementary schools averaged .72, compared to .76 for rural elementary schools. At high schools, the average phi coefficient associated with conducting only five SOMs was .60. Although reliabilities for high schools were somewhat lower than were those from elementary schools, this may be a statistical artifact due to the lower number of high schools in the sample compared to the number of elementary schools. It is recommended that a minimum of five SOMs be conducted at a school to maintain adequate levels of reliability. Additional SOMs can be conducted at a school and tables are provided for educators to balance the cost of conducting additional SOMs with the asymptotic rate of return in increased reliability.

Appendix A

SOM Instrument

Appendix B

Norms for SOM Items

Table 7

Percent of time classroom practices were observed in Elementary Schools

SOM Item	Never	Rarely	Occasionally	Frequently	Extensively	Mean	St Dev
Direct Instruction	1	5	18	43	32	3.01	.877
Team teaching	50	36	11	3	1	.676	.809
Cooperative learning	36	37	19	7	2	1.03	.999
Individual tutoring	43	36	14	6	1	.874	.952
Ability groups	35	25	17	15	8	1.35	1.31
Multi-age grouping	55	20	12	9	4	.874	1.17
Workcenters	32	36	21	9	2	1.13	1.02
Higher level feedback	13	18	24	28	18	2.19	1.28
Integration of subject areas	59	29	8	3	1	.577	.826
Project-based learning	74	20	5	1	1	.329	.620
Higher level questioning	28	35	25	9	3	1.23	1.04
Teacher as Coach	17	27	26	21	10	1.80	1.22
Parent Involvement	76	21	3	1	0	.300	.590
Independent seatwork	3	14	34	35	15	2.45	.993
Experiential, hands-on learning	33	42	19	5	1	1.04	.917
Individual instruction	70	22	6	2	1	.398	.691
Sustained writing	53	37	9	1	0	.584	.709
Sustained reading	39	40	14	6	2	.922	.951
Independent inquiry	82	15	3	1	0	.214	.495
Student discussion	59	25	9	4	3	.651	.975
Computer for instructional delivery	50	35	12	3	1	.702	.843
Technology as a Learning tool	63	27	8	1	1	.489	.724
Performance assessment	68	21	7	4	1	.491	.841
Student self assessment	77	15	6	1	1	.330	.692

Summary Items	Low	Medium	High	Mean	St Dev
Academic focus	2	40	58	2.55	.543
Student attention	4	51	45	2.41	.562

Note. The scale for classroom practices ranged from 0 (never) to 4 (extensively) and the scale for the two summary items ranged from 1 (low) to 3 (high).

Table 8

Percent of time classroom practices were observed in High Schools

SOM Item	Never	Rarely	Occasionally	Frequently	Extensively	Mean	St Dev
Direct Instruction	0	9	22	50	19	2.78	.858
Team teaching	89	11	1	0	0	.120	.342
Cooperative learning	31	47	17	6	0	.978	.847
Individual tutoring	52	28	15	5	0	.724	.894
Ability groups	47	25	8	11	8	1.09	1.33
Multi-age grouping	33	18	22	23	4	1.48	1.28
Workcenters	55	34	10	1	0	.568	.713
Higher level feedback	10	22	32	27	9	2.04	1.12
Integration of subject areas	62	33	4	1	0	.435	.606
Project-based learning	54	34	10	2	0	.599	.742
Higher level questioning	23	52	22	3	1	1.07	.780
Teacher as Coach	13	27	38	19	3	1.72	1.02
Parent Involvement	95	5	0	0	0	.049	.217
Independent seatwork	3	15	35	38	10	2.38	.949
Experiential, hands-on learning	29	45	21	3	1	1.02	.856
Individual instruction	79	20	1	0	0	.225	.445
Sustained writing	64	30	6	1	0	.437	.651
Sustained reading	63	32	4	1	0	.427	.622
Independent inquiry	60	33	7	1	0	.473	.644
Student discussion	50	34	14	3	1	.712	.842
Computer for instructional delivery	50	42	7	1	0	.578	.647
Technology as a Learning tool	32	45	21	2	0	.919	.772
Performance assessment	59	33	7	1	0	.489	.653
Student self assessment	80	18	2	0	0	.222	.466

Summary Items	Low	Medium	High	Mean	St Dev
Academic focus	5	48	47	2.41	.594
Student attention	5	66	29	2.23	.538

Note. The scale for classroom practices ranged from 0 (never) to 4 (extensively) and the scale for the two summary items ranged from 1 (low) to 3 (high).

References

Berliner, D. (1976). Impediments to the study of teacher effectiveness. *Journal of Teacher Education, 27*(1), 5-13.

Brophy, J. & Evertson, C. (1976). *Learning from teaching: A developmental perspective.* Boston: Allyn and Bacon.

Brophy, J., & Good, T. (1986). In Wittrock (Ed.), *Handbook of research on teaching* (pp. 328-375). New York: Macmillan.

Brophey, J., Coulter, C., Crawford, W., Evertson, C., & King, C. (1975). Classroom observation scales: Stability across time and context and relationships with student learning gains. Journal of Educational Psychology, 67, 873-881.

Capie, W. & Ellett, C. (1982). Issues in the measurement of teacher competencies: Validity, reliability, and practicality of Georgia's assessment program (Report No. NP 82-1). Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Cooley, W. & Mao, B. (1980). The sample of classroom time observed. Paper presented at the American Educational Research Association symposium, Reading Instruction in Classrooms for the Learning Disabled.

Dunkin, M., & Biddle, B. (1975). *The study of teaching.* New York: Holt, Rhinehart, & Winston.

Erlich, O. & Shavelson, R. (1978). The search for correlations between measures of teacher behavior and student achievement: Measurement problem, conceptualizaiong problem, or both? *Journal of Educational Measurement, 15*, 77-89.

Evertson, C. & Veldman, D. (1981). Changes over time in process measures of classroom behavior. Journal of Educational Psychology, 73 (2), 156-163.

Flanders, N. (1970). *Analyzing teacher behavior.* Reading, MA:Addisson-Wesley.

Kane, M., & Brennan, R. (1977). The generalizability of class means. *Review of Educational Research 47*, 267-292.

Karweit, N. & Slavin, R. (1982). Time-on-task: Issues of timing, sampling, and definition. *Journal of Educational Psychology, 74*(6), 844-851.

Lewis, E.M., Ross, S.M., & Alberg, M.J. (1999). School Observation Measure: Reliability analysis. Memphis, TN: The University of Memphis, Center for Research in Educational Policy.

Lomax, R. (1982). An application of generalizability theory to observational research. *Journal of Experimental Education*, *51*(1) 22-30.

Marshall, H (1975). Clarification of open education: An analysis of the Walberg & Thomas scales. *Research in Education*.

Marshall, H., Green, J., & Lawrence, M. (1976). Stability of teacher behaviors as measured by a broad range low-inference observational system. Paper presented at the American Educational Research Association (San Francisco, CA).

McGreal, T. (1980). Helping teachers set goals. *Educational Leadership, 37*, 414-420.

Medley, D., & Mitzel, H., (1963). Measuring classroom behavior by systematic observation. In N.L. Gage (Ed.) Handbook of research on teaching. Chicago: Rand McNally.

No Child Left Behind Act of 2001. (H.R. 1), 110 (2002) (enacted).

Padilla, M., Capie, W., Cronin, L. (1986). The infludence of team size and observer-type on the validity and reliability of assessment decisions. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Rosenshine, B., (1970). Evaluation of classroom instruction. *Review of Educational Research, 40*, 279-300.

Ross, S., Alberg, M., Smith, L., Anderson, R., Bol, L., Dietrich, A., Lowther, D., & Phillipsen, L. (2000). Using whole-school restructuring designs to improve educational outcomes. Teaching and Change, 7(2), 111-126.

Ross, S. M., Smith, L. J., Lohr, L., McNelis, M. J., Nunnery, J., & Rich, L. (1991). *Final report: 1991 classroom observation study*. Memphis, TN: The University of Memphis, Center for Research in Educational Policy.

Rothenberg, L, & Hessling, P. (1990). Applying the APE/AERA/NCME "Standards:: Evidence for the validity and reliability of three statewide teaching assessment instruments. Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA: April 16-20).

Rowley, G. (1976). The reliability of observational measures. American Educational Journal, 13, 51-59.

Rowley, G., (1978). The relationship of reliability in classroom research to the amount of observation: An extension of the Spearman-Brown formula. Journal of Educational Measurement, 15, 165-180.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. London: Sage Publications.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist, 44*, 922-932.

Tobin, K. & Capie, W., (1981). Measuring pupil engagement. Paper presented at the annual meeting of the American Educational Research Association. Los Angeles.