DOCUMENT RESUME

ED 335 400 TM 017 062

AUTHOR Teddlie, Charles; And Others
TITLE A Study of the Generalizability of the System for Teaching and Learning Assessment and Review (STAR).
PUB DATE Apr 90
NOTE 34p.; For related papers, see TM 017 059 and TM 017 099.
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150) -- Tests/Evaluation Instruments (160)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Beginning Teachers; Classroom Observation Techniques; *Elementary School Teachers; Elementary Secondary Education; *Evaluation Criteria; Evaluation Methods; Evaluators; *Generalizability Theory; Learning Strategies; *Reliability; *Secondary School Teachers; State Programs; Teacher Certification; *Teacher Evaluation; Teaching Methods; Testing Programs
IDENTIFIERS *System for Teaching Learning Assessment Review; Teacher Performance Appraisal System

ABSTRACT

The results are provided of an initial analysis of the reliability (generalizability) of the System for Teaching and Learning Assessment and Review (STAR) as a comprehensive measure of classroom teaching and learning for making teacher certification decisions. The STAR contains 140 indicators of teacher effectiveness and student learning, which are classified into four performance dimensions (preparation, planning, evaluation; classroom/behavior management; learning enviornment; enhancement of learning) that are operationalized by 23 teaching and learning components (TLCs). The STAR is used to train principals, master teachers, supervisors, college faculty, and other educators to thoroughly assess beginning and experienced teachers' classroom performances for the purpose of renewable professional certification. Study data were collected during the spring of 1989 in 11 schools in an urban school district in southeast Louisiana. Forty-six teachers were assessed on the STAR on 2 occasions by each of the 3 observers for a total of 276 assessments (46 teachers x 6 observations). A four-facet General Purpose Analysis of Variance System was used with the following factors: teachers, assessor type, assessment indicators, and assessment occasions. A generalizability coefficient (GC) was computed for each of the STAR TLCs. The results suggest common perspectives across assessor types as they view classroom teaching and learning over multiple teachers and multiple lessons. The GC for the STAR thinking skills component was the highest of all GCs. The results generally support the STAR's initial reliability. One figure and four data tables are provided. (RLC)

A STUDY OF THE GENERALIZABILITY OF THE SYSTEM FOR TEACHING AND LEARNING ASSESSMENT AND REVIEW (STAR)

Charles Teddlie
Chad Ellett
Nitin Naik

College of Education
113 Peabody Hall
Louisiana State University
Baton Rouge, LA 70803

Presented at the annual meeting of the American Educational Research Association, April 1990, Boston, MA.

April 1990

The System for Teaching and learning Assessment and Review (STAR) has been developed to meet the requirements of two recent pieces of Louisiana legislation: the 1984 Teaching Internship Program law and the 1988 Children First Act, which called for renewable teaching certificates for all state teachers. The STAR is a comprehensive, on-the-job assessment process designed to build on the efforts of other states to identify and assess elements of teaching reflected in the extant process/product literature on effective teaching (Brophy, 1986; Porter & Brophy, 1986) and newer concerns about the assessment of knowledge of content, pedagogy and curriculum (Shulman, 1986).

The STAR has been designed to assess key indicators of teacher effectiveness. An initial assessment framework was developed for the STAR based upon a content synthesis of assessment items derived from 8 other state systems (Ellett, Garland & Logan, 1987; Logan, Garland & Ellett, 1989). This synthesis was considered the "baseline" for the subsequent development of STAR assessment indicators, and several additions have been made to broaden perspectives on a new generation of assessments of teaching and learning (Ellett, 1990). In particular, items have been developed to assess the effective teaching of thinking skills and to assess student learning. Thus, the STAR is being developed in a way that moves the teacher assessment field forward in terms of what is measured within the context of a state mandated teacher evaluation program.

The current version of the STAR (Ellett, Loup & Chauvin, 1989) contains 140 indicators of teacher effectiveness and student learning. These indicators are classified into four Performance Dimensions (Preparation, Planning, Evaluation; Classroom/Behavior Management; Learning Environment; Enhancement of Learning) operationalized by 23 Teaching and Learning Components. The components include concepts such as lesson

initiation, pace, sequence, aids/materials, time management, maintaining appropriate behavior, routines, thinking skills, monitoring learning, informal assessment, etc. The STAR is completed by a three person assessment team for each teacher: the principal, a master teacher, and an "external" evaluator.

This paper reports the results of an initial and continuing series of analyses of the reliability of the STAR as a comprehensive measure of classroom teaching and learning to make certification decisions. The reliability model used reflects a comprehensive data collection system similar to those developed in the past in other states such as Georgia. Past investigations of the reliability of these systems that include the use of multiple data collectors over multiple occasions have proven to be quite promising (Capie, Tobin, Ellett & Johnson, 1981; Capie & Ellett, 1982; Performance Assessment Systems, 1984). The study reported here extends this work, since the STAR has been designed to assess the effectiveness of teacher performance and student learning at the same time.

All analyses were completed using A General Purpose Analysis of Variance System (GENOVA), (Crick & Brennan, 1983). Generalizability theory (Brennan, 1978; Crocker & Algina, 1986; Cronbach, Gleser, Nanda & Rajaratnan, 1972; Medley & Mitzel, 1963) was selected as the method of choice for the analyses. In its derivation from analysis of variance, GENOVA allows for identifying and estimating multiple sources of variation simultaneously. It has the added benefit of providing for the simulation of alternative data collection strategies such as variations in numbers of observers or observation categories. A properly designed study which generates a high generalizability coefficient provides evidence that the assessment system can differentiate subjects (i.e., teachers) in terms of their abilities, while generalizing over assessors (i.e., agreement among principal, master teacher and external),

items (i.e., internal consistency of assessment indicators and components) and assessment occasion (i.e., fall and spring assessments). When coefficients are lower than desired, examination of variance components for facets in the design can suggest where there may be undesirable variation in the data.

The purpose of this paper is to report the results of an initial investigation of the reliability (generalizability) of a statewide, on-the-job assessment system designed for use to renew the professional certificates of all 45,000 teachers in Louisiana. The dependability (for making decisions) of the system will be addressed in future studies.

Data Sources and Methods

Data for this study were collected during the late spring of 1989 in eleven schools in an urban school district in southeast Louisiana. Altogether 46 teachers were assessed on the STAR on two occasions by each of three observer types (principal, master teacher, outside observer). All data were collected confidentially, and no discussion of results with assessed teachers occurred until all six observations were completed and summarized. A total of 276 assessments were completed (46 teachers x 6 observations).

The observers in this study were trained by project staff immediately preceding data collection. All assessors, except for the outside observers, completed an abbreviated 4-5 day training program, but were considered proficient enough to conduct accurate assessments. The outside observers were project staff members, who had not only been trained, but had also provided extensive training to others on the STAR system.

The STAR assessment process required assessors to observe for the full period of a lesson (typically 50-55 minutes) while taking comprehensive notes including periodic estimates of student engagement rates. The observation focus is the total classroom

learning environment, not simply an "evaluation" of the teacher's behavior and performance (Chauvin, Ellett & Loup, 1990). All completed assessments were returned to the project office for keypunching and data analysis.

A four facet GENOVA model was utilized with the following factors: teachers (the only random factor), assessor types, assessment indicators and assessment occasions. The model was fully crossed. Of interest in the analyses was the extent to which the STAR data collection model could differentiate teacher performance on the STAR teaching and learning components and generalize scores over assessor type, assessment indicators, and assessment occasions. A generalizability coefficient was computed for each of the STAR teaching and learning components, since the eventual scoring system will be a criterion-referenced system using a performance standard for each component. Each teaching and learning component is scored by summing a series of dichotomous scores (Acceptable/Unacceptable) for a set of performance indicators defining each component. These indicator scores were summed across each assessors' decisions on each observation to yield a component score for each assessor for each of two occasions.

This procedure yielded scores for seventeen components across three performance dimensions: classroom/behavior management (5 components); learning environment (2 components); and enhancement of learning (10 components). The other performance dimension of preparation, planning, evaluation (which had 6 components) was not analyzed in this study. This dimension is not assessed with direct classroom observations but rather with analysis of a comprehensive unit plan. Results from a separate study of the Spring 1989 data on that dimension will be reported elsewhere.

The results reported in this paper will include: (1) descriptive statistics comparing the percentage of maximum possible score for each component broken down by assessor type and assessment occasion; (2) generalizability coefficients for both indicators and components comparing models with two or three observers; and (3) variance estimates for the components. Figure 1 lists the STAR performance dimensions and components that will be discussed in this paper.

The STAR is based on the assumption that a teaching/learning component is a complex set of interrelated behaviors. Each component is defined by a number of assessment indicators. As noted in Figure 1, the number of assessment indicators per teaching/learning component, for the seventeen components reported here, range from one (student engagement) to fifteen (psychosocial learning environment).

Generalizability theory (Cronbach et al., 1972) was used to plan the various analyses. The procedure used in this study is similar to that described by Capie et al. (1981). Five facets in the analysis design were identified as important sources of variation in the performance data obtained: teachers; assessors; assessor types; occasion of measurement; and assessment indicators. The five-facet design with assessors nested within assessor-types is identical to a four-facet fully crossed design with teachers, assessor-types, occasion of measurement and performance indicators as the sources of variation. As a consequence, the simpler four-facet model was used in all analyses.

For each analysis teachers were treated as facets of differentiation and assessor type, assessment occasion and assessment indicators within teaching/learning components were treated as facets of generalization. A strong case can be made for treating each facet of generalization as fixed in the reliability model. There are only three assessor types involved

in the assessment (principal, master teacher, external assessor), and, although individuals within types do vary, the three types exist as a fixed team for all on-the-job observations. Thus, assessor type was regarded as a fixed facet in the analysis design.

Similarly, the assessment indicators are not random representations of the teaching/learning components. The indicators were constructed to represent the most essential elements of each component. While there are certainly other indicators for each component, they are not considered equal in importance to the set incorporated in the STAR. Likewise, assessment occasions are not randomly selected. Rather, they are special occasions where the teacher endeavors to perform in a "best fashion" that may well be atypical of everyday performances. Consequently, each facet other than teachers was regarded as fixed in the fully crossed design.

Results

Percentages of the maximum possible scores for each of the three observers across performance dimensions and components are shown in Table 1. These percentages suggest great stability across occasions; that is, assessors tended to arrive at similar assessment decisions on the first and second observations.

The three assessor types differed in terms of their assessment decisions. Master teachers judged a higher percentage of the teachers to have satisfactorily attained components than either principals or external observers. Similarly, principals judged the teachers more highly than external assessors. When weighted by the number of indicators, the average percentage of the maximum possible score was 74.5% for master teachers; 65.7% for principals; and 55.8% for external assessors.

There was also considerable variation in the percentage of the maximum scores for individual assessment indicators. In general, higher scores were evident for learning environment components, followed by classroom behavior management components, and then by enhancement of learning components. The two components having the highest percentage of the maximum possible scores were component #23 (oral/written communication) with over 95% attainment, and component #13 (physical learning environment) with almost 90% attainment.

On the other hand, several components received much lower scores:

Component #9 (student engagement) - 45.4%;

Component #10 (managing task-related behavior) - 51.9%;

Component #14 (lesson/activities initiation) - 50.3%;

Component #19 (thinking skills) - 44.1%; and

Component #22 (feedback) - 49.7%.

These were components on which there was the most variation between the master teacher and the external assessor, with the external assessor scores falling in the 30-40% range and the master teacher scores falling in the 50-60% range. The principal scores fell in between these scores.

Generalizability coefficients for two types of models are found in Table 2 (for teaching/learning components) and Table 3 (for assessment indicators). Both models (one for the principal and the external assessor; the other for all three assessors) involve a process that simulates a three-assessor model. The second model adds the effect of the third assessor (master teacher) to that of the first two assessors (principal and external). The addition of the master teacher into the model in this order is important, since it is unclear

whether or not master teachers will be part of the teacher assessment process in Louisiana. The way the current law is written, addition of the master teacher to the assessment model depends upon the availability of funds and results from pilot research with the STAR system.

For all components except oral/written communication, the generalizability coefficients for a model with the principal and the external assessor were .45 or better. These coefficients increased to .50 or better for the model with all three assessors entered. The average generalizability coefficient increased from .61 to .67 when the master teacher assessor was added to the other two assessors. Interestingly, the coefficients for Thinking Skills (Comp. #19) were the highest of all (.73; .81).

The generalizability coefficient for the oral/written communication component is very low. This is probably due to the fact that over all observers and occasions, 96.3% of the assessment decisions were positive. With such ceiling effects, there is no meaningful variability in the data and the resultant generalizability coefficient is lowered.

Generalizability coefficients for 112 performance indicators with either the principal and external assessor entered or for all three assessors entered are shown in Table 3. The data in this table are useful for looking at information on the effects of specific performance indicators as they are added to the component scores (additive mode). Again excluding the oral/written communication indicators, the average generalizability coefficient increases from .41 to .50 when the effects of the master teacher are added to that of the other two assessors. This average .09 increase in the indicator generalizability coefficients is greater than the .06 average increase seen in the results for the generalizability coefficients for STAR components. Typically there is a steady increase in component reliabilities as indicators are added to the analyses, with the largest increases occuring as the number of indicators increase from one to four or five.

Variance component estimates for the 16 STAR components are found in Table 4. These variance component estimates are similar to those reported by Capie et al. (1981) for the Teacher Performance Assessment Instruments (TPAI) used in Georgia for the initial certification of beginning teachers when dichotomous decisions are used, as is the case with the STAR.

Examination of the variance components in Table 4 shows from two to six or seven times as much variation for all variables in the model (TROI) as for variation in teachers (T) alone. The size of the variance components for facets in the design varies considerably from one facet to another and in interactions between facets. These results suggest great stability in the data from one assessment occasion to the next (O), and when occasions are considered in relationship to assessor types (RO), assessment indicators (OI), and assessor types and indicators combined (ROI). These latter interactions are important because they represent variance components that should be low in relationship to others in the analyses. A high variance component for occasion by indicator (OI), for example, would suggest an undesirable interaction between facets in the model and an indication of great instability in assessment decisions across indicators from one occasion to the next. Assessor type by indicator (RI) and assessor type by teacher (TR) variance components are typically much larger than assessor type by occasion (RO) variance components. Considered collectively, the results in Table 4 suggest that facets in the STAR assessment model are behaving in a manner that supports the reliability of the STAR as a complex data collection and measurement system.

Discussion

This report is an initial study of the reliability of the STAR system, based on a pilot study of 46 teachers. Larger studies involving more teachers are currently underway in Louisiana, but the data from this study provide some preliminary information about the STAR. A limit of the present study is that principals and master teachers underwent an abbreviated 4-5 day training program on the system, while ongoing reliability studies require that these assessors undergo the complete 7-day training program.

Looking first at the data on the percentage of the maximum possible scores given by each of the three assessors, some interesting results are found. In general, there appear to be consistent decisions among the three assessor types across the 17 components in Table 1. In fact, if the percentage scores given by the three assessors are correlated, the following results are found: the correlation between the principals' percentage scores and the externals percentage scores across the 17 categories is .97; the correlation between the principals' and the master teachers' scores is .91; and that between the master teachers' and the externals' is .95.

This means that as percentage scores by components increase for one group of assessors, they also increase for the other assessor groups. Thus, assessors' average percentage scores across components in the STAR are highly consistent. For instance, all three assessors judged teachers highly on component 13 (physical learning environment), while judging them relatively low on component 19 (thinking skills). These results suggest common perspectives across assessor types as they view classroom teaching and learning over multiple teachers and multiple lessons.

While all three assessor types agree in terms of the relative percentage of teachers satisfactorily mastering components, there are some differences in their assessments with master teachers giving higher scores than principals, who give higher scores than externals. This may be partly a function of the training that the master teachers and principals received. Since they did not undergo the full 7-day training program, they might not have felt comfortable assessing all the individual performance indicators and may have had a tendency to give higher scores on certain indicators if they were indecisive. Perhaps this "halo" effect did not occur with the externals, who were the most experienced assessors in the state having served as the trainers for the STAR system statewide during the six months preceding the study.

The data from this generalizability study provide a preliminary estimate of the reliability of the STAR as a data collection system. The average generalizability coefficient across 14 of the components with the effect of all three evaluators considered was .67. Given the preliminary nature of this study, a generalizability coefficient of this magnitude seems reasonable. This finding is consistent with those for other on-the-job assessment systems reported elsewhere (Capie, Ellett & Cronin, 1985). Also, the increased reliability associated with the addition of the effect of the master teacher gives some credence to the argument that a third assessor should be part of the STAR assessment model. This is especially the case when one examines the increase of .09 in the average generalizability coefficient for individual assessment indicators when the effect of the master teacher is entered into the analyses. Interestingly, the generalizability coefficients for the STAR Thinking Skills Component were the highest of all coefficients. This finding is important because this is a relatively new assessment area not represented as thoroughly on other on-the-job

assessment systems. Training in the STAR in Thinking Skills assessment is somewhat difficult relative to other assessment areas. The high reliabilities for Thinking Skills reported here suggest that STAR assessors can differentiate teachers' skills to enhance students' abilities to think.

There are two important factors that need to be discussed in considering the preliminary results on the reliability of the STAR reported here. First, these data were collected as part of a research study and when the STAR is fully implemented under "high stakes" conditions, teachers will likely be more knowledgeable and motivated to "try harder". This may make assessors abilities to differentiate performances somewhat more difficult. Even the best trained and most well-intentioned STAR assessor may be somewhat influenced by the "assessment demand characteristics" and become either more "hard-nosed" or generous. These effects have been demonstrated with other systems (Capie, Ellett & Johnson, 1982) with the suggestion that "true" reliabilities need to be established with data resulting from "high stakes" assessments for certification.

Secondly, the use of the external assessor seems imperative in the STAR assessment model to prevent the possibility of "halo" or artificially inflated scores because of the in-building social contexts of assessments for principals and master teachers. The fact that master teachers in this study gave the highest scores suggests that they may need to be more discriminating in their judgements. Or, perhaps, the teachers assessed actually performed better when viewed by their colleagues! Whatever the case, the addition of the master teacher to the STAR assessment team does increase the reliability of the assessment data and including the master teacher may show typically higher scores overall than a model using only the principal and the external.

The results reported in this study are generally supportive of the initial reliability of the STAR. They should be viewed in light of the results of a much larger (n=150 teachers) study currently underway in Louisiana with more highly trained assessors and with a "tightened" version of the STAR being used in the 1989-1990 extended pilot program. When this larger study is completed, performance standards for the STAR and the dependability of eventual certification decisions can be explored.

References

Brennan, R.L. (1978). Extensions of generalizability theory to domain reference testing. Iowa City: American College Testing Program.

Brophy, J. (1986, October). Teacher influences on student achievement. American Psychologist, 1069-1077.

Capie, W. & Ellett, C.D. (1982, March). The effects of assessment demand characteristics on the dependability of teacher performance measures. Paper presented at the annual meeting of the American Educational Research Association, New York.

Capie, W., Ellett, C.D. & Cronin, L. (1985). Assessing meritorious teacher performance: Reliability, decision-making and standards setting procedures. Paper presented at the annual meeting of the American Educational Research Association, Chicago, Illinois.

Capie, W., Ellett, C.D. & Johnson (1982). The effects of assessment demand characteristics on the dependability of teacher performance measures. Paper presented at the annual meeting of the American Educational Research Association, New York, N.Y.

Capie, W., Tobin, K., Ellett, C.D. & Johnson, C. (1981, April). The dependability of job performance rating scales for making classification decisions. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

Crick, J.E., & Brennan, R.L. (1983). GENOVA: A general purpose analyses of variance system. Iowa City: American College Testing Program.

Crocker, L. & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart & Winston.

Cronbach, L.J., Gieser, G.C., Nanda, H. & Rajaratnan, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley.

Ellett, C.D., Loup, K. & Chauvin, S. (1989). System for Teaching and learning Assessment and Review (STAR). Statewide Teaching Internship and Teacher Evaluation Form. Baton Rouge, LA: College of Education, Louisiana State University, Louisiana Department of Education.

Ellett, C.D., Garland, J. & Logan, C. (1987). Content classification, synthesis and verification of eight large-scale teacher performance assessment instruments. Research report, Teaching Internship Project, Baton Rouge, LA: College of Education, Louisiana State University.

Ellett, C.D. (1990). An new generation of classroom-based assessments of teaching and learning: Concepts, issues and controversies from pilots of the Louisiana STAR. College of Education, Louisiana State University, Baton Rouge, LA.

Logan, C., Garland, J. & Ellett, C.D. (1989). Large-scale teacher performance assessment instruments: A synthesis of what they measure and a national survey of their influence on the preparation of teachers. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, California.

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

Porter, A.C. & Brophy, J. (1988). Synthesis of research on good teaching: Insights from the work of the institute for research on teaching. Educational Leadership, 74-85.

Performance Assessment Systems, Inc. (1984). A study of the generalizability of the Teacher Assessment and Development System (TADS) MTP form. Athens, GA: Author.

Shulman, L. (1986). Those who understand: Knowledge growth in teaching. Educational Researcher, 15 (2), 4-14.

PERFORMANCE DIMENSION II : CLASSROOM BEHAVIOR MANAGEMENT (30)

TEACHING/LEARNING COMPONENTS

A.	Time (8)	Component # 7
B.	Classroom Routines (4)	Component # 8
C.	Student Engagement (1)	Component # 9
D.	Managing Task-Related Behavior (7)	Component #10
E.	Monitoring/Maintaining Student Behavior (10)	Component #11

PERFORMANCE DIMENSION III: LEARNING ENVIRONMENT (20)

TEACHING/LEARNING COMPONENTS

A.	Psychosocial Learning Environment (15)	Component #12
B.	Physical Learning Environment (5)	Component #13

PERFORMANCE DIMENSION IV: ENHANCEMENT OF LEARNING (68)

TEACHING/LEARNING COMPONENTS

A.	Lesson/Activities Initiation (10)	Component #14
B.	Teaching Methods (5)	Component #15
C.	Sequence/Pace (5)	Component #16
D.	Aids and Materials (10)	Component #17
E.	Content Accuracy/Emphasis (8)	Component #18
F.	Thinking Skills (11)	Component #19
G.	Clarification (5)	Component #20
H.	Monitoring Learning Activities/ Informal Assessment (6)	Component #21
I.	Feedback (4)	Component #22
J.	Oral/Written Communication (4)	Component #23

NOTE: The numbers in parentheses note the number of assessment indicators comprising the three performance dimensions and seventeen teaching/learning components.

Figure 1

STAR Performance Dimension and Teaching/Learning Components
Used in Generalizability Study

Table 1

Percent of Maximum Possible Scores Given by Each of Three STAR Assessors Across Performance Dimensions and Components

	Principal		Master Teacher		External Assessor	
	First Observation	Second Observation	First Observation	Second Observation	First Observation	Second Observation
Dimen 2	69.3	69.6	72.1	71.4	58.2	57.9
Comp. # 7	76.4	71.8	79.9	77.6	70.5	69.6
Comp. # 8	80.8	83.0	79.8	81.5	76.5	70.8
Comp. # 9	48.9	46.8	63.8	55.3	25.5	31.9
Comp. #10	51.0	52.9	61.1	59.0	43.1	44.4
Comp. #11	73.8	76.6	71.9	72.8	54.7	55.3
Dimen 3	75.1	74.8	82.8	84.8	71.6	67.8
Comp. #12	70.1	69.9	80.1	81.3	66.3	63.4
Comp. #13	90.2	89.4	90.6	95.4	87.6	80.8
Dimen 4	61.5	61.3	73.7	72.3	50.6	50.9
Comp. #14	53.2	51.1	65.9	60.6	33.8	37.0
Comp. #15	71.4	71.4	84.6	84.2	69.0	63.8
Comp. #16	66.8	65.6	77.0	78.8	55.8	56.2
Comp. #17	71.9	68.3	86.4	86.2	65.7	68.9
Comp. #18	66.3	64.6	79.0	78.3	55.0	57.8
Comp. #19	45.5	51.1	54.0	54.4	30.7	29.2
Comp. #20	66.4	64.6	75.8	71.0	57.8	59.2
Comp. #21	53.2	52.5	71.0	66.3	34.0	36.2
Comp. #22	45.3	46.8	65.5	66.5	41.0	33.0
Comp. #23	95.3	96.8	96.8	97.8	96.8	94.3

Table 2

Generalizability Coefficients for the STAR Teaching/Learning Components

Teaching/ Learning Component	G-Coefficient: Principal and External Assessor	G-Coefficient Principal, External Assessor and Master Teacher
# 7 Time	.598	.643
# 8 Classroom Routines	.525	.577
#10 Managing Task-Related Behavior	.645	.700
#11 Monitoring/Maintaining Student Behavior	.723	.775
#12 Psychosocial Learning Environment	.726	.789
#13 Physical Learning Environment	.631	.695
#14 Lessons/Activities Initiation	.664	.722
#15 Teaching Methods	.577	.630
#16 Sequence/Pace	.521	.576
#17 Aids and Materials	.614	.682
#18 Content Accuracy/ Emphasis	.660	.728
#19 Thinking Skills	.732	.807
#20 Clarification	.447	.497
#21 Monitoring Learning Activities/Informal Assessment	.596	.651
#22 Feedback	.625	.691
#23 Oral/Written Communication	.130	.147

NOTE: Both models presented here simulate a three observer model. The second model adds the effect of the third observer (master teacher) to that of the first two observers (principal and external assessor).

Table 3

Generalizability Coefficients for the STAR Performance Indicators (Additive Model)

Performance Indicator	Principal and External Assessor	Principal, External Assessor and Master Teacher	Performance Indicator	Principal and External Assessor	Principal, External Assessor and Master Teacher
Comp. #7 - Time			Comp. #10 - Managing Task-Related Behavior		
Indicator 1	.31	.38	Indicator 1	.41	.51
2	.39	.48	2	.49	.60
3	.43	.53	3	.52	.64
4	.45	.56	4	.53	.66
5	.47	.58	5	.54	.67
6	.48	.59	6	.55	.68
7	.49	.60	7	.56	.68
8	.49	.61			
Comp. #8 - Classroom Routine					
Indicator 1	.29	.38			
2	.36	.46			
3	.39	.50			
4	.41	.52			

Table 3 - Continued

Generalizability Coefficients for the STAR Performance Indicators (Additive Model)

Performance Indicator Master Teacher	Principal and External	Principal, External Assessor and Assessor	Performance Indicator Master Teacher	Principal and External	Principal, External Assessor and Assessor
Comp. #11 - Managing/Maintaining Student Behavior			Comp #12 - Psychosocial Learning Environment		
Indicator 1	.42	.51	Indicator 1	.38	.46
2	.50	.61	2	.47	.57
3	.53	.64	3	.51	.62
4	.55	.67	4	.54	.64
5	.56	.68	5	.55	.66
6	.57	.69	6	.56	.67
7	.57	.69	7	.57	.68
8	.58	.70	8	.58	.69
9	.58	.70	9	.58	.69
10	.59	.71	10	.59	.70
			11	.59	.70
			12	.59	.70

Table 3 - Continued

Generalizability Coefficients for the STAR Performance Indicators (Additive Model)

Performance Indicator Master Teacher		Principal and External	Principal, External Assessor and Assessor	Performance Indicator Master Teacher		Principal and External	Principal, External Assessor and Assessor
Comp. #13 - Physical Learning Environment				Comp. #14 - Lesson/Activities Initiation			
Indicator	1	.42	.49	Indicator	1	.32	.39
	2	.53	.47		2	.39	.47
	3	.58	.69		3	.43	.53
	4	.61	.73		4	.45	.55
	5	.63	.75		5	.46	.57
					6	.47	.58
					7	.48	.59
					8	.48	.59
					9	.49	.60
					10	.49	.60

Table 3 - Continued

Generalizability Coefficients for the STAR Performance Indicators (Additive Model)

Performance Indicator Master Teacher	Principal and External	Principal, External Assessor and Assessor	Performance Indicator Master Teacher	Principal and External	Principal, External Assessor and Assessor
Comp. #15 - Teaching Methods			Comp. #17 - Aids and Materials		
Indicator 1	.36	.45	Indicator 1	.31	.37
2	.44	.54	2	.38	.46
3	.48	.59	3	.42	.50
4	.50	.61	4	.43	.52
5	.51	.63	5	.44	.54
Comp. 16 - Sequence/Pace			6	.45	.55
Indicator 1	.29	.34	7	.46	.56
2	.36	.44	8	.46	.56
3	.39	.48	9	.47	.57
4	.41	.51	10	.47	.57
5	.42	.52			

Table 3 - Continued

Generalizability Coefficients for the STAR Performance Indicators (Additive Model)

Performance Indicator Master Teacher	Principal and External	Principal, External Assessor and Assessor	Performance Indicator Master Teacher	Principal and External	Principal, External Assessor and Assessor
Comp. #18 - Content Accuracy/Emphasis			**Comp. #19 - Thinking Skills**		
Indicator 1	.34	.40	Indicator 1	.38	.45
2	.43	.51	2	.47	.56
3	.47	.57	3	.51	.61
4	.50	.60	4	.54	.64
5	.51	.62	5	.55	.66
6	.52	.63	6	.56	.67
7	.53	.64	7	.57	.68
8	.54	.65	8	.58	.69
			9	.58	.69
			10	.59	.70
			11	.59	.70

Table 3 - Continued

Generalizability Coefficients for the STAR Performance Indicators (Additive Model)

Performance Indicator Master Teacher	Principal and External	Principal, External Assessor and Assessor
Comp. #20 - Clarification		
Indicator 1	.30	.39
2	.36	.47
3	.38	.50
4	.39	.51
5	.40	.52
Comp. #21 - Monitoring Learning Activities/Informal Assessment		
Indicator 1	.35	.43
2	.42	.52
3	.46	.56
4	.47	.58
5	.49	.60
6	.50	.61
Comp. #22 - Feedback		
Indicator 1	.39	.47
2	.46	.56
3	.49	.61
4	.51	.63
Comp. #23 - Oral/Written Communication		
Indicator 1	.04	.04
2	.05	.05
3	.06	.06
4	.06	.06

NOTE: Both models presented here simulate a three observer model. The second model adds the effect of the third observer (master teacher) to that of the first two observers (principal and external evaluator).

Table 4

Variance Component Estimates for Sixteen STAR Components
Based on Acceptable/Unacceptable Decisions

Source of Variation	#7	#8	#10	#11	#12	#13	#14	#15	#16	#17	#18	#19	#20	#21	#22	#23
Teacher (T)	.0199	.0264	.0594	.0533	.0262	.0120	.0319	.0312	.0160	.0162	.0156	.0299	.0318	.0390	.0378	.0000
Assessor Type (R)	.0014Q	.0011Q	.0053Q	.0111Q	.0055Q	.0009Q	.0187Q	.0078Q	.0113Q	.0096Q	.0119Q	.0147Q	.0039Q	.0271Q	.0201Q	.0000Q
Occasion (O)	.0001	.0000	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
Indicator (I)	.0480Q	.0112Q	.0045Q	.0136Q	.0187Q	.0026Q	.0181Q	.0127Q	.0641Q	.0260Q	.0667Q	.0051Q	.0047Q	.0094Q	.0117Q	.0001Q
TR	.0056	.0180	.0367	.0255	.0207	.0097	.0025	.0023	.0182	.0266	.0188	.0388	.0288	.0259	.0548	.0028
TO	.0062	.0081	.0166	.0090	.0059	.0042	0..85	.0128	.0084	.0075	.0043	.0061	.0165	.0126	.0181	.0065
TI	.0191	.0072	.0199	.0161	.0204	.0189	.0181	.0180	.0161	.0145	.0221	.0288	.0107	.0136	.0178	.0000
RO	.0000	.0002	.0000	.0000	.0000	.0013	.0005	.0000	.0000	.0001	.0000	.0003	.0000	.0000	.0006	.0000
RI	.0037Q	.0013Q	.0001Q	.0036Q	.0029Q	.0006Q	.0073Q	.0017Q	.0082Q	.0026Q	.0050Q	.0050Q	.0003Q	.0069Q	.0040Q	.0000Q
OI	.0000	.0000	.0000	.0001	.0000	.0000	.0004	.0001	.0002	.0000	.0000	.0015	.0000	.0000	.0001	.0000
TRO	.0154	.0357	.0549	.0287	.0143	.0157	.0222	.0399	.0282	.0250	.0138	.0188	.0942	.0440	.0411	.0178
TRI	.0313	.0221	.0261	.0251	.0459	.0366	.0399	.0268	.0245	.0233	.0384	.0617	.0188	.0441	.0297	.0000
TOI	.0278	.0341	.0292	.0273	.0213	.0094	.0373	.0269	.0356	.0291	.0264	.0336	.0278	.0357	.0320	.0081
ROI	.0001	.0000	.0008	.0004	.0015	.0000	.0010	.0000	.0000	.0000	.0004	.0002	.0000	.0000	.0042	.0000
TROI	.0817	.0933	.0870	.0738	.0731	.0356	.1208	.0767	.0934	.0795	.0796	.0988	.0922	.1032	.1046	.0226