# DOCUMENT RESUME

ED 335 408 TM 017 092

AUTHOR Ellett, Chad D.; And Others

TITLE The Effects of "High Stakes" Certification Demands on the Generalizability and Dependability of a Classroom-Based Teacher Assessment System.

PUB DATE Apr 91

NOTE 33p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1991).

PUB TYPE Statistical Data (110) -- Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS Classroom Observation Techniques; *Elementary School Teachers; *Evaluation Criteria; Evaluators; Generalizability Theory; Learning Strategies; *Reliability; *Secondary School Teachers; State Programs; Statistical Data; Teacher Certification; *Teacher Evaluation; Teaching Methods; Testing Programs

IDENTIFIERS Educational Indicators; *Louisiana; *System for Teaching Learning Assessment Review LA; Teacher Performance Appraisal System

ABSTRACT

This paper presents the results of ongoing analyses of the reliability 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. Focus was on the effects of high stakes assessment conditions for certification on the generalizability evidence for the STAR when compared to more normal research conditions. The STAR contains 140 teacher effectiveness and student learning indicators, which are classified into four performance dimensions that are operationalized by 23 teaching and learning components (TLCs). Data were collected during the spring of 1990 using 144 teachers in 30 elementary and secondary schools in southeastern Louisiana school districts. In all, 864 observations were completed (144 teachers x 6 observations). Also, data from STAR assessments (16,524 observations) for 2,754 teachers and 2 random samples of about 100 teachers nested within this group were used. A four-facet (teachers, assessor type, occasion of measurement, and assessment indicators) General Purpose Analysis of Variance System was used. A generalizability coefficient (GC) was computed for each STAR TLC. Descriptive statistics for various STAR TLCs, GCs for components comparing models with two or three observers, GCs collected under research versus certification conditions, and variance component estimates for the TLCs for a statewide certification data file are reported. Assessment demand characteristics under certification conditions can alter the statistical reliabilities of classroom-based teacher assessment systems. These conditions typically skew performance distributions, resulting in decreased statistical reliabilities. The STAR's reliability seems best supported by coefficients established under research conditions with no high stakes certification demands. Nine data tables and a 16-item list of references are included. (RLC)

# The Effects of "High Stakes" Certification Demands on the Generalizability and Dependability of a Classroom-Based Teacher Assessment System

Chad D. Ellett, Charles Teddlie and Nitin Naik

College of Education

Louisiana State University

Paper presented at the annual meeting
of the ' merican Educational Research Association,
Chicago, Illinois

April, 1991

# The Effects of "High Stakes" Certification Demands on the Generalizability and Dependability of a Classroom-Based Teacher Assessment System

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, more constructivist concerns about the assessment of knowledge of content, pedagogy and curriculum (Shulman, 1986).

The STAR has been designed to assess key indicators of effective teaching and learning. 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.

Two versions of the STAR assessment framework were used in the studies reported here. The pilot version (Ellett, Loup & Chauvin, 1989) which was comprised of 140 assessment indicators of the quality of teaching and learning and a 1990 certification version consisting of 117 assessment indicators. STAR assessment 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 a continuing series of analyses of the reliability of the STAR as a comprehensive measure of classroom teaching and learning. The reliability model used reflects a comprehensive data collection system similar to those developed in the past in other states such as Georgia and Texas. 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; Teddlie, Ellett & Naik, 1990). The studies reported here extend this earlier work, since the STAR has been designed to assess the effectiveness of teaching and to make "in situ" inferences about 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 and extends classical approaches to estimating reliability of measurements. 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 within the context reported here which generates a high generalizability coefficient provides evidence that the assessment system can differentiate subjects (i.e., teachers/classrooms) in terms of their teaching and learning qualities, while generalizing scores over assessors (i.e., agreement among principal, master teacher and external), items (i.e., internal consistency of assessment indicators and components) and assessment occasions (i.e., fall and spring assessments). When coefficients are lower than desired, examination of variance components for facets in the analysis design can suggest where there may be undesirable variations in the data and the assessment model.

## Purpose

The purpose of this paper is to report the results of a series of continuing investigations of the reliability (generalizability) of the STAR assessment model as it has been developed and initially implemented statewide for making teacher certification decisions in Louisiana. Of particular interest are the effects of "high stakes" assessment conditions for certification on the generalizability evidence for this data collection system when compared to more "normal" research conditions.

## Data Sources and Methods

Data for this study were collected during the late spring of 1990 in a research study using 144 teachers in 30 elementary and secondary schools in a mixed set of urban and rural districts in southeast Louisiana. In addition, data were available for analysis from a complete set of

STAR certification assessments from a large sample of 2754 teachers and two random samples of approximately 100 teachers nested within this larger group of certification assessments.

The STAR is administered in fall and spring by three assessor "types" (principal, master teacher, outside assessor/state employee). In the smaller (1990) research study, each assessor type completed two observations of a particular teacher over a three to four week period to simulate the process as it was being designed for the initial certification year (1990-1991). Multiple groups of assessors were assigned to the 144 teachers assessed in the research study. 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 864 observations were completed in the research study (144 teachers x 6 observations) The larger data set (n=2754 complete assessments or 16,524 observations) was made available through a statewide file of STAR data derived from initial implementation of the STAR for purposes of professional, renewable certification in Louisiana.

The observers/assessors in this study were all fully trained and certified in the STAR assessment process through a comprehensive 7-day assessor certification program. The STAR assessment process requires assessors to observe for the full period of a lesson (typically 50-55 minutes) while taking comprehensive notes including periodic estimates of the quantity and quality of student engagement in learning tasks. The STAR observation focus is the total classroom learning environment, not simply an "evaluation" of the teacher's behavior and performance (Ellett, 1990).

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 teaching and learning quality as assessed by the STAR teaching and learning components and generalize scores over assessor types, assessment indicators, and assessment occasions. A generalizability coefficient was computed for each of the STAR teaching and learning components, since the eventual STAR decisionmaking framework is a criterion-referenced system using a performance standard for each component. Each teaching and learning component is scored by summing a series of dichotomous decisions (Acceptable/Unacceptable) for a set of assessment indicators defining each component. These assessment indicator decisions are then summed for each assessor to yield a component score for each assessor for each of two assessment occasions for each teacher assessed.

In the studies reported here, this procedure yielded scores for sixteen STAR teaching and learning components in the research study and fifteen teaching and learning components in the "certification" study. One component ("Pace") was deleted in the 1990 revision of the STAR document. These component scores are distributed across three STAR performance dimensions: Classroom/Behavior Management, Learning Environment and Enhancement of Learning. The fourth STAR performance dimension of Preparation, Planning and Evaluation was not included in this study. This dimension is not assessed with direct classroom observations but rather with analysis of a comprehensive unit plan via semi-structured interviews.

Results reported in this paper include: (1) descriptive statistics for various STAR teaching, and learning components, (2) generalizability coefficients for components comparing models with two or three observers; and (3) generalizabilty coefficients collected under research versus certification conditions; (4) variance component estimates for the teaching and learning components for a statewide certification data file.

The STAR is based on the assumption that a teaching/learning component represents a complex set of interrelated behaviors. Each component is defined by a number of assessment indicators. The number of assessment indicators per teaching/learning component, for the sixteen and fifteen components reported here ranged from three (physical learning environment) to eleven (thinking skills).

## GENOVA Design

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 assessment 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. However, the arrangement of times for fall and spring occasions for observations for certification is somewhat random across teachers. In addition, occasions for the research study data set were arranged, but were rather random across teachers in the study relative to all of the lessons teachers might teach. Therefore, the occasion facet was treated as random in the GENOVA model used in the analyses reported here.

Within the GENOVA design, three sets of generalizability (G) coefficients were generated with the data sets available for the various STAR teaching and learning components: 1) coefficients for assessments from a spring, 1990 research study that simulated the STAR assessment process; 2) coefficients for a large set (n=2754) of complete STAR assessments (all six observations, three fall and three spring) collected during the initial year of statewide program implementation in 1990-1991; and 3) two random samples of complete STAR assessments derived from the larger sample of certification assessments. For these samples, separate G coefficients were computed with various combinations of assessor types to examine the effects of adding or deleting a third assessor to the assessment model and STAR assessment process.

In a GENOVA study, two kinds of coefficients are typically generated: 1) Generalizability (G) coefficients which are a useful index of the ability of the data collection system to differentiate subjects in terms of their relative rank order; and 2) "phi" coefficients which are useful in making absolute decisions relative to a designated standard. Phi coefficients are typically somewhat lower in magnitude than G coefficients because they are computed with a larger error term derived from all main and interaction effects in the GENOVA model. In this paper, only G coefficients will be reported.

## Results

### Research Context

Table 1 presents a summary of Generalizability (G) coefficients for STAR teaching and learning components derived from complete STAR assessments completed on a sample of elementary and secondary teachers participating in an assessment "simulation" study during the late spring of 1990. Coefficients are shown for the principal and external/state assessor "types" without the master teacher and for the complete team of three assessor types with the addition of the master teacher. For the two-member team, G coefficients ranged from .22 (Time) to .62 (Teaching Methods and Learning Tasks). For this team, 11 of 16 coefficients approached or exceeded .40 and six coefficients approached or exceeded .50. The table mean coefficient for the two-member team was .42. For the three-member team, adding the master teacher assessor, STAR teaching and learning component G coefficients ranged from a low of .29 (Time) to a high of .70 (Teaching Methods and Learning Tasks). For this sample of 144 teachers, three-member team coefficients were, in all cases somewhat higher than for the two-member team analysis. For the three-member team, 14 of 16 coefficients approached or exceeded .40 and 11 approached or exceeded .50. The mean G coefficient for the three-member team was .51. These results indicate the "value-addedness" of the master teacher as part of the STAR assessment team. However, the coefficients presented in Table 1 are somewhat lower than those previously reported for the STAR (Teddlie, Ellett & Naik, 1990). In this prior study, the G coefficients for STAR teaching and learning components for the three-member team were somewhat more "robust" and ranged from .50 (Clarification) to .81 (Thinking Skills), with a mean coefficient of .67 (see Table 9).

## Certification Context

Table 2 presents a summary of STAR teaching and learning component G coefficients derived from an analysis of a large sample of 2754 complete STAR assessments collected under conditions of making professional, renewable certification decisions in Louisiana during the 1990-1991 school year. The results shown in Table 2 are considerably lower than those shown in Table 1 for STAR data collected under research ("low stakes") conditions. For the two-member team, G coefficients ranged from .16 (Clarification) to .32 (Thinking Skills). Only five of 15 component coefficients approached or exceeded .30. The mean G coefficient across all components for the two-member team was .23.

For the three-member assessment team (adding the master teacher), G coefficients under "high stakes" conditions shown in Table 2 ranged from low of .18 (Clarification) to .35 (Thinking Skills). For the three-member team, only 7 of 15 coefficients approached or exceeded .30. The mean coefficient across all 15 STAR teaching and learning components for the three-member team for this large sample of assessments was .25.

Table 3 provides G coefficients for two and thee-member STAR assessment teams for a random sample of 103 teachers selected from the larger data set of 2754 complete certification assessments. Table 4 provides the same kind of results for a second random sample of 101 complete STAR assessments from the larger data set collected under certification conditions. In comparing results in the two tables, several interesting findings emerge. First, in all instances, the G coefficients for the second random sample (Table 4) are higher than those for the first random sample (Table 3). The mean, two-person team G coefficient for the first random sample was .27 (Table 3) and this same result for the second random sample was .46. For the three-member teams, the mean coefficient for the first random sample (Table 3) was .29 and for the second random sample (Table 4) was .49.

Secondly, some interesting anomalies are evident in comparing the results in Tables 3 and 4. For example, the G coefficient for the first random sample of teachers (Table 3) for Oral/Written Communication for the two-member team was .10. This same coefficient for the second random sample (Table 4) was .71. A similar comparison for the STAR teaching and learning component of Physical Learning Environment (#13) shows coefficients of .19 for teacher sample number one (Table 3) and .60 for teacher sample number two (Table 4).

Thirdly, G coefficients and the table mean coefficient for the second random sample (Table 4) more closely approximate those established under research conditions with the STAR during the spring of 1990 (Table 1). However, these coefficients as well are lower than those previously reported for the STAR teaching and learning components (Teddlie, et al., 1990) (Table 9).

Table 5 presents a summary of G coefficients for selected STAR teaching and learning components for separate sets of fall, 1990 and spring, 1991 assessments (three observations in each data set). Results are shown in Table 5 for both a two- and a three-member STAR assessment teams, again adding in the master teacher. These results were generated foı the large sample (n=2754) of teachers assessed under certification conditions. The GENOVA design used to generate these coefficients was only a three-facet design since there is no occasion (fall and spring) effect in the data analysis model. Therefore, the composition of error terms and the ratios between various variance components in the model are altered considerably from the four-facet model that includes the occasion facet.

The results in Table 5 are interesting when making two comparisons. First, they are typically higher than the G coefficients reported for the entire data set (all six observations) for the four-facet model shown in Table 2. Secondly, spring coefficients are slightly higher than fall coefficients for every STAR component.

Table 6 presents a summary of G coefficients for a two-member team consisting of the external/state assessor and the master teacher assessor. Thus, for these analyses, the principal assessor type was removed. The results in Table 6 show a set of G coefficients that are higher in all cases than the coefficients in Table 2 computed on the same sample of 2754 complete STAR assessments using all three assessor types. In comparing the results in Tables 6 and 2, the G coefficients for STAR components 7 through 14 are of particular interest and implications of these findings will be discussed below.

Table 7 presents a summary of variance components for each main and interaction effect in the GENOVA model for the analysis of the complete certification data set of 2754 assessments. The variance components in this table can be used to examine "lawful" and "unlawful" effects in the STAR assessment model given the assumptions on which it is based and the design of the STAR as a data collection system. Variance components are included for each of the four-facets in the GENOVA model and their various interaction terms for each of the STAR teaching and learning components.

It is beyond the scope of this paper to discuss more than the "highlights" of the findings in Table 7. However, inspection of these results shows some interesting findings:

1. in comparing the variance components for the four facets in the design, the teacher (T) facet of differentiation is typically larger than the other three facets of generalization (R, O and I).
2. the variance components associated with assessor "type" are rather small
3. some of the "unlawful" interaction terms (e.g., RO and ROI) are rather small relative to other variance components
4. the assessor type by indicator (RI) variance components are relatively small (an "unlawful" interaction) while the teacher by indicator (TI) and teacher by occasion (TO) variance components are much higher ("lawful" interactions)

A summary of descriptive statistics for each STAR teaching and learning component is for the certification sample of 2754 STAR assessments for fall, 1990 and spring, 1991 assessment occasions is presented in Table 8. Included in the table is an index of the mean expressed as a percentage of the maximum possible score. The number of assessment indicators varies from one STAR component to the next and this index allows for a comparison of component scores in terms of "mastery" levels. The "possible" scores for components are determined by multiplying the number of assessment indicators for a component X 3 observations (principal, master teacher external assessor/state employee). Only the standard deviations for various components having the same possible score are directly comparable.

For the fall, 1990 assessment, the %Max. scores ranged from a low of 62.67% (Thinking Skills) to a high of 99.00 (Oral/Written Communication) with scores in the 80s and 90s most typical. In comparing fall, 1990 scores to spring, 1991 scores, the %Max. increased for all 15 STAR teaching and learning components and the component standard deviations decreased in magnitude for all 15 components. The increases in "mastery" scores and decreases in score variability reflect two concerns: 1) improvements from fall to spring; and/or 2) possible artificial score "inflation."

## Discussion, Implications and Conclusions

The purpose of this study was to compare reliability (generalizability) coefficients for decisionmaking components of a comprehensive, classroom-based system design to assess teaching and learning computed under research ("low stakes") and teacher certification ("high stakes") conditions. A variety of analyses were completed on data collected under these conditions using the System for Teaching and learning Assessment and Review (STAR) (Ellett, Loup & Chauvin, 1989). Data were analyzed within parameters specified by a General Purpose Analysis of Variance System (GENOVA) (Crick & Brennan, 1983) in a complex, four-facet analysis design. In addition to comparing generalizability (G) coefficients computed under, research and certification conditions, separate analyses were completed by deleting and adding STAR assessor "types" (principal, master teacher, external assessor/state employee) and for fall and spring sets of assessment data collected under certification conditions.

The results reported here generally support the earlier findings of Capie & Ellett (1982) that assessment "demand characteristics" under conditions of certification can alter the statistical reliabilities of classroom-based, teacher assessment systems. These conditions typically skew performance distributions (higher scores when compared to research conditions) with the result that statistical reliabilities decrease. This general finding from the Capie & Ellett (1982) study and the research reported here raises some interesting issues about reliability characteristics and their relationship to making important decisions such as teacher certification. For example, several sets of G coefficients for the STAR have been presented in this paper and these vary considerably in value from one research study to the next and under research vs. certification conditions. Which coefficients are the most important to consider in making certification decisions and examining the "dependability" of a measurement system like the STAR to make classification decisions?

On the one hand, G coefficients established under research ("low stakes") conditions with the STAR are appealing because they better mirror variations in the "everyday" effectiveness of teaching and learning. Thus, they may more adequately represent the "real" statistical reliability of the STAR as a data collection system. On the other hand, reliabilities established under actual certification conditions may attest more to the actual dependability ("trustworthiness") of the system for making certification decisions and the probability of making "false positive" and "false negative" decisions relative to established performance standards. Thus, the reliability of the measurement and data collection system alone seems best supported by coefficients established under research conditions where the "high stakes" demands of certification are not present.

Such demands in the STAR system are many, including the fact that teachers have access to the identity of assessors' scores which may foster artificial score inflation or "pollution." The school principal, for example, may have given a teacher five to ten years of positive evaluations on the local district evaluation form and during the certification assessment, there may exist considerable social and psychological "press" to make assessment scores conform with this past "halo." In the STAR certification data set used in this study, considerable evidence for this "press" exists since principals scored teachers considerably higher than the other two assessor types, and principals' assessments "passed" teachers at twice the rate of the other assessors. This score inflation also seems evident in the TRI and TRO variance components (Table 7) in the certification data set when compared to variance components for T (teachers) when these same comparisons are made with results established under research conditions previously reported in Teddlie, Ellett & Naik, (1990).

A principal's presence in the classroom might also serve to change the observation environment in ways that produce unwanted "noise" in an observation system. For example, removal of the principal from the analyses reported in Table 6 for STAR components 7-14 show a considerable increase in the magnitude of the G coefficient when these values are compared to those reflected in Table 2 for the three-person team. Interestingly, these STAR components focus on Time and Classroom/Behavior Management concerns.

Comparisons of G coefficients established under both research and certification conditions for two- and three-member assessment teams demonstrate the value-addedness of the third team member. Under both conditions, G coefficients were higher for the three-member than for the two-member team. This might be expected given the demonstrated relationship between classical approaches to test length and increases in reliability. However, it should be remembered that the magnitude of G coefficients in complex measurement and data collection systems like the STAR depends on more than test length alone. Interestingly, in this regard (and given the discussion above about school principals' assessments), the magnitude of G coefficients in the certification data set actually increased when deleting principals' scores and recomputing STAR component G coefficients with a two-member team (master teacher and external assessor/state employee). Similar analyses (removing the principals' scores) under research conditions with the STAR show decreases in the magnitude of G coefficients, not increases.

The two random samples selected from the larger (n=2574) data set of certification assessments were somewhat disparate in the value of STAR component G coefficients (Tables 3 and 4). In addition, both were somewhat higher than coefficients established for the entire data set (Table 2). This seems rather surprising. However, it should be remembered that according to sampling theory, repeated random samples should generate a distribution of G coefficient values varying considerably in magnitude. Thus, using a random sampling procedure on one or two samples from a large data set like that used in this study (n=2574) to establish reliability estimates may generate inflated (or deflated) reliability estimates. For example, a third sample taken from this large data set (not reported here) yielded G coefficients for STAR Components in the range of .60 to .85 that were considerably higher than those reported for the entire certification data set. Thus, some caution seems in order if simple random selection procedures are used to make reliability estimates. While useful, it appears that a variety of such samples would need to be taken to generate a reasonably stable estimate of "true" reliabilities.

Separate G coefficients were computed on the certification data set for fall assessments (three observations for each teacher and for spring assessments (three observations for each teacher) for selected STAR teaching and learning components (Table 5). These coefficients were somewhat higher than those computed for the complete (fall and spring) set of assessments (six observations). Of course, the GENOVA design used in these analyses deletes the "occasion" facet and the design becomes a three-facet rather than a four-facet design. These findings suggest that there is a considerable occasion effect in the STAR assessment model that reduces statistical reliability. Given the increased magnitude and decreased variability of STAR component scores from fall to spring (Table 8), it seems understandable that interaction terms of facets in the GENOVA design with variance due to assessment occasion would serve to decrease the magnitude of various G coefficients. This, the finding that fall and spring G coefficients were somewhat higher than those for the complete (fall and spring) set of assessments is not surprising.

The above finding should be viewed with some caution since policymakers and others might use this statistical information as a basis for reducing the number of assessments needed to make certification decisions. One might, for example, try to make the argument that demonstrated reliabilities are higher with three observations than with six observations with the STAR. Therefore, the number of required number of observations should be reduced to three, as the argument is presented. This argument seems supported from analyses of this large data set collected under certification conditions.

However, this argument may be misleading. For example, the results of analyses of STAR data collected under research conditions in a previous study (Teddlie, Ellett, & Naik,

1990) and this study do not support this conclusion. Table 9, for example, shows G coefficients for STAR teaching and learning components established under research conditions for a sample of 66 classrooms in the spring of 1989 for a complete set of six assessments. These coefficients are the highest to date for the STAR components and are considerably higher than those repcıted for the certification data set used in this study. The G coefficients reported in Table 9 were computed on a data set collected by STAR assessors that were, perhaps the most highly trained (8 1/2 vs 7 days of training) of any used in other reliability studies with the STAR. Thus, the "expertise" of STAR assessors may be an important variable in adequately differentiating teachers in terms of various STAR assessment indicators.

Of course, if one followed the simple statistical argument presented above for making a decision about changing the current STAR data collection model, the reliability results presented here would suggest that the two-member STAR assessment team, excluding the school principal is a better model than the three-member assessment team. This model, however, cannot currently be put forward because the principal is mandated as part of the assessment team by current state law.

In comparing separate fall and spring results to combined results, one also needs to appreciate the effects of the "occasion" facet described above and the inflation of spring scores when compared to fall scores, particularly by school principals. In the certification assessment data set analyzed in this study, the occasior ˙ 'ect, combined with inflated scores and reduced, variation of scores in the spring assessments, apparently reduced the ability of the STAR system to differentiate teachers in terms of STAR component scores and to generalized these scores over assessment occasions, assessor types and assessment indicators.

Clearly, in deciding upon a data collection design with a complex system like the STAR, a host of issues other than statistical reliability alone need to be considered. Thus, demonstrating adequate statistical reliability is a necessary condition for documenting the psychometric quality of asssessment systems like the STAR. And, as is clearly shown here, such reliabilities can vary considerably from one study/analysis to the next given the great variety of assessment demand characteristics and parameters that can surround data collection (Capie & Ellett, 1982). However, establishing adequate statistical reliability is not sufficient to establish the professional "credibility" of a system like the STAR nor to fulfill the variety of purposes for which the system was originally developed.

## 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., 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 ot 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., Gleser, 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. Cnicago: 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.

Teddlie, C., Ellett, C.D., & Naik, N. (1990). A study of the generalizability of the System for Teaching and learning Assessment and Review (STAR). Paper presented at the annual meeting of the American Educational Research Association, Boston, Massachusetts.

Table 1

Generalizability Coefficients for the STAR Teaching and Learning Components

| Teaching and Learning Components | | G-Coefficient Principal and External Assessor | G-Coefficient Principal, External Assessor and Master Teacher |
|---|---|---|---|
| **PERFORMANCE DIMENSION II: CLASSROOM AND BEHAVIOR MANAGEMENT** | | | |
| A. | Time | 0.223 | 0.292 |
| B. | Classroom Routines | 0.441 | 0.540 |
| D. | Managing Task-Related-Behavior | 0.595 | 0.683 |
| E. | Monitoring and Maintaining Student Behavior | 0.561 | 0.655 |
| **PERFORMANCE DIMENSION III: LEARNING ENVIRONMENT** | | | |
| A. | Psychsocial | 0.461 | 0.557 |
| B. | Physical | 0.30 | 0.391 |
| **PERFORMANCE DIMENSION IV: ENHANCEMENT OF LEARNING** | | | |
| A. | Lesson and Activities Initiation | 0.397 | 0.497 |
| B. | Teaching Methods and Learning Tasks | 0.616 | 0.702 |
| C. | Aids and Materials | 0.386 | 0.463 |
| D. | Content Accuracy and Emphasis | 0.383 | 0.483 |
| E. | Thinking Skills | 0.433 | 0.526 |
| F. | Clarification | 0.327 | 0.419 |
| G. | Pace | 0.268 | 0.355 |
| H. | Monitoring Learning Tasks and Informal Assessment | 0.560 | 0.647 |
| I. | Feedback | 0.370 | 0.462 |
| J. | Oral and Written Communication | 0.340 | 0.435 |

Table 2

Generalizability Coefficients for the STAR Teaching/Learning Components
(n=2754)

| Teaching/ Learning Component | G-Coefficient: Principal and External Assessor | G-Coefficient Principal, External Assessor and Master Teacher |
|---|---|---|
| # 7 Time | .2102 | .2315 |
| # 8 Classroom Routines | .1679 | .1857 |
| #10 Managing Task-Related Behavior | .2731 | .2986 |
| #11 Monitoring/Maintaining Student Behavior | .2553 | .2802 |
| #12 Psychosocial Learning Environment | .2600 | .2847 |
| #13 Physical Learning Environment | .2652 | .2915 |
| #14 Lessons/Activities Initiation | .2621 | .2875 |
| #15 Teaching Methods | .2282 | .2508 |
| #16 Aids and Materials | .1949 | .2153 |
| #17 Content Accuracy/ Emphasis | .2057 | .2263 |
| #18 Thinking Skills | .3230 | .3515 |
| #19 Clarification | .1631 | .1807 |
| #20 Monitoring Learning Activities/Informal Assessment | .2334 | .2565 |
| #21 Feedback | .1788 | .1979 |
| #22 Oral/Written Communication | .2382 | .2615 |

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 Teaching/Learning Components (n=103)

| Teaching/ Learning Component | G-Coefficient: Principal and External Assessor | G-Coefficient Ptincipal, External Assessor and Master Teacher |
|---|---|---|
| # 7 Time | .3202 | .3510 |
| # 8 Classroom Routines | .2039 | .2253 |
| #10 Managing Task-Related Behavior | .2659 | .2902 |
| #11 Monitoring/Maintaining Student Behavior | .3976 | .4357 |
| #12 Psychosocial Learning Environment | .4172 | .4536 |
| #13 Physical Learning Environment | .1881 | .2079 |
| #14 Lessons/Activities Initiation | .3078 | .3354 |
| #15 Teaching Methods | .2544 | .2794 |
| #16 Aids and Materials | .3702 | .4013 |
| #17 Content Accuracy/ Emphasis | .2640 | .2877 |
| #18 Thinking Skills | .3117 | .3408 |
| #19 Clarification | .1584 | .1762 |
| #20 Monitoring Learning Activities/Informal Assessment | .2632 | .2889 |
| #21 Feedback | .1578 | .1749 |
| #22 Oral/Written Communication | .1024 | .1147 |

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 4

Generalizability Coefficients for the STAR Teaching/Learning Components
(n=101)

| Teaching/ Learning Component | G-Coefficient: Principal and External Assessor | G-Coefficient Principal, External Assessor and Master Teacher |
|---|---|---|
| # 7 Time | .5218 | .5589 |
| # 8 Classroom Routines | .5414 | .5721 |
| #10 Managing Task-Related Behavior | .5023 | .5369 |
| #11 Monitoring/Maintaining Student Behavior | .4232 | .4538 |
| #12 Psychosocial Learning Environment | .5602 | .5908 |
| #13 Physical Learning Environment | .5965 | .6273 |
| #14 Lessons/Activities Initiation | .4631 | .4959 |
| #15 Teaching Methods | .4134 | .4434 |
| #16 Aids and Materials | .4230 | .4535 |
| #17 Content Accuracy/ Emphasis | .3708 | .4000 |
| #18 Thinking Skills | .4224 | .4548 |
| #19 Clarification | .3534 | .3827 |
| #20 Monitoring Learning Activities/Informal Assessment | .3669 | .3964 |
| #21 Feedback | .2827 | .3088 |
| #22 Oral/Written Communication | .7147 | .7400 |

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 5

Generalizability Coefficients for the STAR Teaching/Learning Components (n=2754)

| Teaching/ Learning Component | G-Coefficient: Principal and External Assessor | G-Coefficient Principal, External Assessor and Master Teacher |
|---|---|---|
| Spring | | |
| #10 Managing Task-Related Behavior | .4442 | .5452 |
| #11 Monitoring/Maintaining Student Behavior | .3867 | .4861 |
| #12 Psychosocial Learning Environment | .4177 | .5182 |
| #13 Physical Learning Environment | .3857 | .4849 |
| #18 Thinking Skills | .5031 | .6030 |
| Fall | | |
| #10 Managing Task-Related Behavior | .4791 | .5798 |
| #11 Monitoring/Maintaining Student Behavior | .4151 | .5156 |
| #12 Psychosocial Learning Environment | .4918 | .5921 |
| #13 Physical Learning Environment | .4699 | .5707 |
| #18 Thinking Skills | .5534 | .6502 |

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 6

Generalizability Coefficients for the STAR Teaching/Learning Components
(n=2754)

| Teaching/ Learning Component | Master Teacher and External Assessor |
|---|---|
| # 7 Time | .3782 |
| # 8 Classroom Routines | .2949 |
| #10 Managing Task-Related Behavior | .4586 |
| #11 Monitoring/Maintaining Student Behavior | .4536 |
| #12 Psychosocial Learning Environment | .4230 |
| #13 Physical Learning Environment | .4728 |
| #14 Lessons/Activities Initiation | .4335 |
| #15 Teaching Methods | .3741 |
| #16 Aids and Materials | .3404 |
| #17 Content Accuracy/ Emphasis | .3188 |
| #18 Thinking Skills | .3050 |
| #19 Clarification | .2340 |
| #20 Monitoring Learning Activities/Informal Assessment | .3700 |
| #21 Feedback | .3209 |
| #22 Oral/Written Communication | .3615 |

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 7

Variance Component Estimates for Sixteen STAR Components
Based on Acceptable/Unacceptable Decisions
(n=2754)

| Source of Variation | #7 | #8 | #10 | #11 | #12 | #13 | #14 | #15 | #16 | #17 | #18 | #19 | #20 | #21 | #22 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Teacher (T) | .0028 | .0016 | .0082 | .0089 | .0024 | .0016 | .0089 | .0040 | .0035 | .0064 | .0147 | .0041 | .0057 | .0090 | .0008 |
| Assessor Type (R) | .0000Q | .0001Q | .0001Q | .0000Q | .0005Q | .0000Q | .0000Q | .0004Q | .0000Q | .0017Q | .0000Q | .0000Q | .0006Q | .0000Q | .0000Q |
| Occasion (O) | .0002 | .0000 | .0002 | .0002 | .0001 | .0000 | .0007 | .0001 | .0000 | .0002 | .0016 | .0000 | .0002 | .0005 | .0000 |
| Indicator (I) | .0110Q | .0005Q | .0008Q | .0139Q | .0023Q | .0003Q | .0121Q | .0024Q | .0007Q | .0000Q | .0120Q | .0009Q | .0020Q | .0025Q | .0000Q |
| TR | .0000 | .0000 | .0000 | .0000 | .0000 | .0002 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 |
| TO | .0029 | .0025 | .0063 | .0075 | .0018 | .0016 | .0057 | .0036 | .0042 | .0055 | .0072 | .0060 | .0046 | .0105 | .0007 |
| TI | .0062 | .0013 | .0030 | .0103 | .0043 | .0028 | .0124 | .0055 | .0018 | .0000 | .0146 | .0021 | .0040 | .0051 | .0007 |
| RO | .0015 | .0001 | .0012 | .0077 | .0004 | .0001 | .0078 | .0005 | .0008 | .0004 | .0131 | .0009 | .0012 | .0049 | .0000 |
| RI | .0000Q | .0000Q | .0000Q | .0000Q | .0027Q | .0000Q | .0000Q | .0002Q | .0000Q | .0000Q | .0000Q | .0000Q | .0004Q | .0000 | .0000Q |
| OI | .0003 | .0000 | .0000 | .0001 | .0001 | .0000 | .0002 | .0001 | .0000 | .0062 | .0001 | .0000 | .0002 | .0000 | .0000 |
| TRO | .0116 | .0084 | .0258 | .0336 | .0074 | .0061 | .0279 | .0142 | .0171 | .0195 | .0352 | .0248 | .0193 | .0444 | .0028 |
| TRI | .0000 | .0004 | .0000 | .0000 | .0000 | .0005 | .0000 | .0000 | .0000 | .0019 | .0000 | .0000 | .0000 | .0000 | .0000 |
| TOI | .0144 | .0632 | .0118 | .0159 | .0105 | .0039 | .0311 | .0151 | .0084 | .0325 | .0405 | .0089 | .0170 | .0248 | .0019 |
| ROI | .0040 | .0002 | .0001 | .0067 | .0007 | .0001 | .0042 | .0004 | .0002 | .0186 | .0047 | .0006 | .0004 | .0018 | .0000 |
| TROI | .0512 | .0195 | .0404 | .0641 | .0386 | .0147 | .1127 | .0531 | .0299 | .0921 | .1453 | .0333 | .0602 | .0882 | .0067 |

*Q = Quadratic Form

Table 8

Descriptive Summaries for Various Componets
(n=2754)

| Teaching/Learning Component | | | FALL | | | SPRING | | |
|---|---|---|---|---|---|---|---|---|
| | | Max. Poss. | Mean | % Max | S.D. | Mean | % Max. | S.D. |
| # 7 | Time | 18 | 16.17 | 89.83 | 1.66 | 17.00 | 95.00 | 1.13 |
| # 8 | Classroom Routines | 12 | 11.52 | 96.00 | 0.97 | 11.83 | 98.58 | 0.56 |
| #10 | Managing Task-Related Behavior | 18 | 16.16 | 89.78 | 2.81 | 17.26 | 95.89 | 1.60 |
| #11 | Monitoring/Maintaining Student Behavior | 18 | 15.01 | 83.39 | 2.85 | 16.23 | 90.17 | 2.10 |
| #12 | Psychosocial Learning Environment | 30 | 27.73 | 92.43 | 2.46 | 29.24 | 97.47 | 1.45 |
| #13 | Physical Learning Environment | 9 | 8.76 | 97.33 | 0.63 | 8.89 | 98.78 | 0.45 |
| #14 | Lessons/Activities Initiation | 24 | 18.20 | 75.83 | 3.56 | 20.87 | 86.96 | 2.53 |
| #15 | Teaching Methods | 18 | 16.31 | 90.61 | 1.89 | 17.18 | 95.44 | 1.31 |
| #16 | Aids and Materials | 18 | 16.79 | 93.28 | 2.03 | 17.57 | 97.61 | 1.14 |
| #17 | Content Accuracy/ Emphasis | 18 | 14.50 | 80.56 | 2.47 | 16.38 | 91.00 | 1.67 |
| #18 | Thinking Skills | 33 | 20.68 | 62.67 | 6.04 | 26.71 | 80.94 | 4.12 |
| #19 | Clarification | 12 | 11.14 | 92.83 | 1.48 | 11.61 | 96.75 | 0.93 |
| #20 | Monitoring Learning Activities/Informal Assessment | 20 | 15.89 | 88.28 | 2.30 | 17.17 | 95.39 | 1.42 |
| #21 | Feedback | 12 | 9.77 | 81.42 | 2.09 | 11.01 | 91.75 | 1.30 |
| #22 | Oral/Written Communication | 12 | 11.88 | 99.00 | 0.49 | 11.93 | 99.42 | 0.39 |

Table 9
Generalizability Coefficients for the STAR Teaching/Learning Components for Spring 1989 Research Study
(n=66)

| 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).