DOCUMENT RESUME

ED 335 404 TM 017 082

AUTHOR Lofton, Glenda G.; And Others
TITLE Results of Iterative Standards-Setting Procedures for a Performance-Based System for Renewable Certification.
PUB DATE Apr 91
NOTE 57p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1991).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Academic Standards; *Elementary School Teachers; *Evaluation Criteria; Evaluation Methods; Evaluators; Learning Strategies; *Secondary School Teachers; State Programs; Teacher Certification; *Teacher Evaluation; Teaching Methods; Testing Programs; Workshops
IDENTIFIERS Iterative Methods; Louisiana; Performance Based Certification; *Standard Setting; *System for Teaching Learning Assessment Review LA

ABSTRACT

This report presents the results of an initial, iterative performance standards-setting (SS) task of a comprehensive on-the-job statewide teacher assessment system--the System for Teaching and Learning Assessment and Review (STAR). The 1990-91 STAR assesses and makes inferences about the quality of teaching and learning on sets of assessment indicators defining 21 teaching and learning components (TLCs). It is used to train principals, master teachers, supervisors, college faculty, and other educators to complete thorough assessments of beginning and experienced teachers' classroom performances for the purpose of obtaining renewable professional certification (RPC). The STAR is organized into four performance dimensions defined by subelements/TLCs. In the spring of 1990, a panel of 47 Louisiana educators participated in a 3.5-day workshop in which they used the results of STAR pilot research studies in 1988-90 as critical information for making initial STAR performance standards recommendations (PSRs) for the Louisiana Teaching Internship and the Statewide Teacher Evaluation Programs. The workshop served as a forum for the presentation/discussion of critical professional and program policy and implementation issues pertaining to PSRs. In the fall of 1990, a final recommendation was made relative to an appropriate performance standard for each test or variable under consideration. A set of iterative procedures was completed serially for each STAR TLC. Three PSRs for RPC were made for each STAR TLC. A 21-item list of references, 9 data tables, and 2 graphs are included. A summary of STAR data analysis results for 5,720 Louisiana teachers using 108 indicators is provided. (RLC)

Results of Iterative Standards-Setting Procedures for a Performance-Based System for Renewable Certification

Glenda G. Lofton
Louisiana State University

Chad D. Ellett
Louisiana State University

Sheila W. Chauvin
Louisiana State University

Karen Loup
Louisiana State University

and

Nitin Naik
Louisiana State University

Paper Presented to Division D
at the meeting of the
American Educational Research Association
Chicago, Il
April, 1991

Results of Iterative Standards-Setting Procedures for a Performance-Based System for Renewable Certification

Overview

The Louisiana System for Teaching and learning Assessment and Review (STAR) (Ellett, Loup & Chauvin, 1990) has been developed and piloted throughout Louisiana for the Louisiana Department of Education during the past two years to meet requirements of the Louisiana Teaching Internship Law (1984) and the Children First Act (1988). Following the lead of many other states, Louisiana is implementing an induction process for new teachers that includes comprehensive, classroom-based assessments to be used as a basis for structuring support programs for continuing professional development. The ultimate goal of the Louisiana Teaching Internship Program is to provide new teachers with successful experiences and support to assist them in their professional development during the early years of employment.

The Children First Act provides for on-the-job assessments of all experienced teachers in Louisiana for the purpose of obtaining a professional, renewable certificate with a maximum validity period of five years. This Act mandates that all Louisiana teachers (excluding interns) be evaluated with a standard system for the purpose of validating and/or renewing the professional teaching credential. The STAR and the accompanying set of assessment procedures will be used to make these certification decisions. Star assessment data will also be used to provide comprehensive, diagnostic information to teachers about the quality and effectiveness of teaching and learning to assist them in their continuing professional development. The STAR is being implemented statewide for the first time in Louisiana during the 1990-1991 school year.

Because of the extensive involvement of educators at all levels in developing and implementing these new assessment programs, the STAR has implications that go beyond meeting the requirements of new Louisiana laws. For example, the STAR has been used as a framework for examining selected elements of teacher education programs and as a basis for inservice education and staff development. The STAR has also proven useful in conducting classroom-based research on effective teaching and learning and in understanding elements of teachers' professional experiences, expertise, and perspectives on reflective practice (Ellett, 1990a).

The STAR is a comprehensive, on-the-job assessment instrument and set of procedures designed to make inferences about the quality of teaching and learning in Louisiana's classrooms. The STAR is grounded in some fifteen years of research and development in designing and implementing large-scale teacher evaluation systems in other states (Ellett, Garland & Logan, 1987). However, the STAR document and the assessment processes extend prior generations of teacher evaluation instruments to include important concerns with classroom context variables and student learning as well (Ellett, 1990a).

A variety of ongoing research and development studies designed to investigate the validity, reliability and implementability of the STAR have been completed (Ellett, 1990b; LaMaster, Tobin & Bowen, 1990; Teddlie, Ellett & Naik, 1990; Chauvin, 1990; Loup, Ellett & Chauvin, 1990; Ellett, Loup, Chauvin & Naik, 1990; Ellett, Chauvin, Loup, & Naik, 1990), or are in process. The results of these studies are encouraging and typically they support the STAR as a professionally endorsed, valid and reasonably reliable system designed to fulfill the purposes of the Louisiana Teaching Internship and Statewide Teacher Evaluation efforts.

A very important research and development concern with the STAR is establishing performance standards for making teacher certification decisions. This paper describes the results of an initial STAR standards-setting task completed by a select group of Louisiana educators during the late spring of 1990. It also includes a brief review of subsequent events which culminated in a set of final performance standards for the STAR in the fall of 1990. Issues, procedures, and "cautions" relative to standards-setting are discussed.

Perspectives on Standards-Setting

Setting performance standards for systems like the STAR is not an easy task. It requires that educators consider many different kinds of information as they proceed and that they are reasonably aware of the consequences of their professional judgements relative to the performance standards they recommend. Ultimately, setting standards is a "human" concern because the task requires that professional judgements be made at many points in the process. As the saying goes, "computers don't set standards...humans do." The key question becomes how do we systematize a process for making informed professional judgements that is scientifically based, replicable and results in standards that are reasonably valid and reliable (1985 Standards for Educational and Psychological Testing)? Jaeger (1990) poses a number of issues to be considered: number, qualification, selection and training of persons to recommend and set standards, what they will do and how data will be collected, analyzed and reported.

While the standard-setting literature includes recommendations on some of these issues, most have been made based on paper-and-pencil tests rather than on classroom observation procedures. Many are based on logic rather than empirical findings. Selecting individuals that are: 1) highly informed ("experts"); 2) professionally committed to take the task seriously; and

3) that represent a variety of "stakeholder" perspectives, is a primary concern. Even though standards setting is ultimately a human concern, providing standards setters with different kinds of information is also critical to setting standards that are reasonable, equitable and professionally credible. Therefore, as standards are set, understanding various kinds of information pertaining to: 1) assessment processes and procedures; 2) program implementation policies and practices; 3) data aggregation rules for making decisions; and 4) results of research studies, is also an important concern.

The 1990-1991 STAR is designed to assess and make inferences about the quality of teaching and learning on sets of assessment indicators defining 21 Teaching and Learning Components. The assessment indicators and components comprising the STAR have been professionally endorsed by teachers and other selected educators throughout Louisiana in two separate research studies, as "essential" for both certification and the enhancement of student learning. The STAR is more than a single "test," and it requires consideration of multiple performance standards for multiple, essential elements of effective teaching and learning. Thus, standards-setting considerations and activities with a system like the STAR are much more complex than those for more traditional evaluation instruments used for teacher certification that require that single (whole score) standards be set (e.g., a single score for the National Teacher Examination).

The STAR has been piloted statewide in Louisiana for the past two years (1988-1990). Pilot activities during 1989-1990 included certification of the principal and a master teacher as STAR assessors in every public school building in Louisiana. In addition, a variety of other educators (e.g., college faculty, assistant principals, parish supervisors, LDE personnel) participated in the STAR professional development program to certify assessors. As part of pilot activities during

1989-1990, a large number of STAR assessments (some 6,000) were completed and these assessments involved teachers, students and classrooms in every parish in Louisiana. Thus, there was a considerable amount of information available about characteristics of effective teaching and learning in Louisiana's classrooms. This information was analyzed in a variety of ways and the ensuing results were used by Louisiana educators as they considered an initial set of STAR performance standards. However, at the time initial standards were set, no information about performance on the STAR was available that had been collected under the "real" and "high stakes" conditions of certification. These data have only become available as the STAR is implemented for certification statewide for the first time during the 1990-1991 school year.

The results of the initial standards-setting task which is the primary focus of this report were to be considered only tentative "benchmarks" because STAR performance distributions are likely to change under "real" assessment conditions pertaining to certification. It was intended that initial performance standards be revisited when additional STAR data was collected during the fall of 1990 and the spring of 1991. With assessment systems similar to the STAR, changes in performance distributions are known to affect the dependability of certification decisions and to change misclassification probabilities from those established under "pilot" conditions (Capie & Ellett, 1982). Typically, such performance distributions shift toward somewhat higher scores when assessment data are collected under "real" conditions. Therefore, the initial performance standards were set with the idea that they should not be considered final by any means. Instead they represented a set of reasonable "benchmarks" for teachers to consider as they began the STAR assessment process for renewable certification during the fall of 1990.

The Initial STAR Standards-Setting Task

An initial standards-setting workshop with Louisiana educators to recommend initial performance expectations for the STAR was held in June, 1990 in New Orleans, Louisiana. The purpose of this workshop was to provide a highly informed ("expert") group of Louisiana educators with the results of STAR pilot research studies (1988-1990) to be used as critical information for making initial STAR performance standards recommendations for the Louisiana Teaching Internship and the Statewide Teacher Evaluation Programs. In addition, the workshop served as a forum for the presentation and discussion of critical professional and program policy and implementation issues that pertained to standards recommendations.

Standards-Setting Panelists

Consistent with recommendations of Hambleton (1978) and Shepherd (1980) on the use of several types of judges, a panel of 47 educators from various regions of Louisiana was nominated by LTIP and LTEP regional, STAR professional development program coordinators giving consideration to two essential concerns: 1) knowledge and expertise in the STAR and the LTIP and LTEP programs; and 2) a reasonable balance among panel members relative to position of employment, ethnicity, gender and other key factors. In selecting panel members, an attempt was made to assure that the majority of panelists were regular classroom teachers. All panelists nominated/selected had extensive preparation as STAR assessors and many had served during the 1989-1990 pilot as program assistants in the STAR professional development program to certify assessors.

Table 1 presents a summary of demographic information on the standards-setting panelists that participated in the initial workshop activities. (Table 1 and subsequent tables are in

Appendix A.) As can be seen in the table, 67.4% of the panelists were female, and 39.1% were black. The majority of panelists were regular classroom teachers (56.5%) and approximately 17.5% were school administrators. Slightly less than one fourth of the panelists (23.9%) were college faculty, parish supervisors and LDE personnel. The majority of panelists represented elementary school levels (41.9%) and over half (56.5%) possessed plus thirty, specialist or doctorate degrees. The average years of experience as an educator for these panelists was 19 years, with almost all of this experience in public schools. The average years of teaching experience for the teacher panelists was 18.28 years. Four of these panelists were members of the LDE performance standards, policy recommendations committee. Four of these panelists were LSU LTIP and LTEP regional coordinators of the professional development program to certify STAR assessors. Thus, these panelists represented a highly experienced, educated and reasonably balanced group of "stakeholders" with considerable knowledge of the STAR assessment document, assessment procedures, legal bases of LTIP and LTEP, and existing program implementation procedures and policies.

The LTIP and LTEP Project Director and three LSU project coordinators organized and served as leaders for the standards-setting workshop. The outside consultant for the workshop design was Dr. Richard Jaeger, College of Education, University of North Carolina at Greensboro.

Standards-Settings Task(s)

The standards-setting process, adapted from the work of Jaeger (1990), was an "iterative" one that occurred over three and one-half days of intensive workshop activity. It included providing panelists with normative information as recommended by Hambleton and Eignor (1980), Jaeger

(1982, 1989) and Shepard (1980), and allowing judges to discuss and reconsider their recommendation (Jaeger, 1978, Brandstalter, Davis, and Stocker-Kreichgauer, 1982). In this type of procedure, panelists are provided with successively more and more critical information for discussion and individual performance standards recommendations are made. Data are then tabulated and summarized, and group discussion of results, concerns and pertinent issues follows. Additional recommendations are then made by each panelist and these are again tabulated, summarized and shared with the group for additional discussion and so on. At some point, a final recommendation is then made relative to an appropriate performance standard for each "test" or variable under consideration. A set of iterative procedures was completed serially for each STAR Teaching and Learning Component.

In this workshop, panelists were first provided with an extensive review of the STAR document to clarify understandings of assessment indicators and Teaching and Learning Components since there was considerable variation in the time of the year in which each had completed the STAR professional development program to certify assessors. This review required almost one full day. During the first part of the second day of the workshop, panelists received a lecture/discussion on issues pertaining to the reliability and validity of the STAR using pilot research results, and examples of the relationship between the reliability of the STAR and the "dependability" of the STAR for making classification decisions for certification. This latter information pointed out the relationship between the reliabilities of various STAR Teaching and Learning Components and the possibilities of making "false positive" and "false negative" decisions. The remaining two workshop days were spent in small and large group activities reviewing STAR research data and iteratively making recommendations for performance

standards and discussing results and pertinent issues.

At the end of the last day, panelists made a final recommendation for a performance standard for each STAR Teaching and Learning Component. They also made a "voice vote" recommendation for a performance expectation for intern teachers and a "one-shot" recommendation for a "superior" performance standard for each STAR Teaching and Learning Component for teachers who might be considered for the Louisiana Model Career Options Program as specified in the Children First Act. As a final activity, panelists were asked to carefully review the assessment indicators for each STAR Teaching and Learning Component and to make recommendations about any assessment indicator that they believed to be so "essential" that being judged as "consistently unacceptable" should result in an "Unacceptable" decision for the entire component.

Each panelist was provided with a workshop manual that included workshop objectives, rules and schedules, response forms for recording judgements, participant evaluation forms for various workshop activities and for recording concerns and STAR research results information. The key research information to be considered by each panelist in making recommendations was as follows:

1. descriptive statistical summaries of STAR assessment indicator and Teaching and Learning Component data derived from 1989-1990 STAR assessments completed in 5,473 classrooms/lessons representing every parish and virtually every public school building in Louisiana

These summaries showed the percentages of "Acceptable" and "Unacceptable" performance judgements compiled by Louisiana educators during the 1989-1990 pilot year for elementary,

secondary and total classroom groups. These data were discussed as representative of "routine, daily performances" in Louisiana's classrooms and cautions were emphasized about predicted shifts in performance distributions under "high stakes" assessment conditions for certification during the fall of 1990. This large number of assessments had been completed by principals, master teachers and other educators, drawn from every parish in Louisiana, as part of STAR assessor certification requirements.

2. graphs depicting the frequency and cumulative percentages of "Acceptable" decisions for assessment indicators and "Mastery" scores for each STAR Teaching and Learning Component

These graphs provided panelists with information about various percentages of "Acceptable" decisions for various numbers (sets) of indicators from the pilot STAR assessments. They also depicted the percentages of teachers with "mastery" scores within various, possible score ranges. Considered collectively, these graphs provided panelists with some information about the possible percentages of teachers that would "pass" or "fail" at various "cut points."

3. summaries of generalizability (reliability) coefficients for STAR Teaching and Learning Components from the 1988-1989 and 1989-1990 pilot research studies

These coefficients provided panelists with some information about how much they could "trust" their performance standards recommendations in view of making potential "false positive" and "false negative" decisions. To simplify matters, panelists were informed that "all other factors considered, the higher the 'G' coefficient, the greater is the 'trust' that can be placed in a recommended performance standard."

Sample copies of descriptive statistical summaries for STAR Teaching and Learning

Components and assessment indicators, percentages and cumulative frequency graphs, and generalizability coefficients provided to panelists are shown in Appendix B. A set of information like the set shown in Appendix B was provided for each panelist for each Teaching and Learning Component in STAR Performance Dimensions II (Classroom and Behavior Management), III (Learning Environment) and IV (Enhancement of Learning). Similar information did not exist for STAR Performance Dimension I (Preparation, Planning and Evaluation). However, this dimension was thoroughly studied and discussed by panelists in view of the results of two, small-scale, qualitative studies of teachers' preparation and utilization of Comprehensive Unit Plans (CUPS) as part of assessments during the 1989-1990 pilot year.

STAR Data Matrices and Standards-Setting Judgements

The STAR is a "criterion-referenced" assessment framework (Ellett, 1990a) and each Teaching and Learning Component is considered an "essential" element of effective teaching and learning and each has been professionally verified by teachers and other Louisiana educators as an "essential" element for professional, renewable certification. Thus, a performance standard needs to be recommended for each component.

The number of assessment indicators comprising each STAR component varies, though the number of assessments (6)(three assessors X 2 occasions; fall and spring) does not. As a result of a complete assessment for a year, a matrix of "1's" and "0's" will be generated for each assessment component. An "Acceptable" decision is recorded as a "1" and an "Unacceptable" decision is recorded as a "0." Each of the assessment indicator decisions is considered a sample of the effectiveness of teaching and learning relative to the larger, "key ideas" reflected in a particular component. Of course, the distribution of "1's" and "0's" for a particular STAR

Teaching and Learning Component will predictably be somewhat different from one observation to the next and from fall to spring assessments, for an individual teacher and across teachers.

By way of example, the 1989-1990 STAR Teaching and Learning Component of "Time" was defined by 8 assessment indicators. With this structure, a matrix of 48 assessment decisions would be generated for this Teaching and Learning Component (3 assessors X 8 assessment indicators X 2 occasions; fall and spring assessments) during an assessment year. Other STAR components have different numbers of assessment indicators and thus, the size of the assessment matrix for a particular component varies from one component to the next. Sample data matrices for a STAR Teaching and Learning Component with 8 assessment indicators are included in Appendix C to illustrate hypothetical variation in "Acceptable" and "Unacceptable" performances relative to assessment indicators from one STAR observation and assessment occasion to the next for a single teacher. It should be noted that while the patterns of assessment decisions are somewhat different in the two matrices, the percentages of "Acceptable" decisions ("1's" in the matrices) are the same. In viewing this matrix, it is obvious that a wide variety of patterns of "1's" and "0's" are possible across various STAR Teaching and Learning Components from fall to spring and across various teachers, observation occasions and assessment contexts. This assessment and decision-making structure creates flexibility in the STAR that is needed to accommodate a wide variety of assessment contexts.

Given this structure, and the need to have a common "metric" for standards setting, panelists were asked to make a recommendation for each STAR Teaching and Learning Component of the "percentage of Acceptable decisions that a teacher should be credited with in the complete assessment process in order to meet the minimum standard for professional, renewable

certification in Louisiana." These recommended percentages for each of the STAR Teaching and Learning Components then, represent an initial set of "minimum performance expectations" for Louisiana teachers for obtaining the professional, renewable teaching certificate.

Results and Discussion

A variety of data were available as panelist's made their recommendations from one iteration of judgements to the next. Three recommendations for a performance standard for professional, renewable certification were made for each STAR Teaching and Learning Component: 1) an initial recommendation after studying pertinent research findings and assessment indicators comprising a particular component; 2) a second recommendation after considerable discussion of the first recommendation with other panelists in small groups; and 3) a final recommendation after sharing the results of the second recommendation with the entire group of panelists.

Considering the number of assessment indicators comprising a particular STAR component in question, the research information provided, other panelists' recommendations and pertinent discussion, each panelist was asked to recommend a final performance standard for each component. The recommendation for each component translated into a numerical percentage of the total matrix possibilities that should be "1's" if a teacher is to meet the minimum standard (performance expectation) for professional, renewable certification in Louisiana. After each round of recommendations, panelists' percentages were rank ordered and arithmetical means and standard deviations were computed.

Table 2 presents a summary of means and standard deviations of percentages of "Acceptable" decisions recommended as a performance standard for professional, renewable certification in Louisiana for each STAR Teaching and Learning Component for first, second, and final rounds

of recommendations. The mean scores represent arithmetical averages computed across each round of recommendations of all panelists. The standard deviations (S.D.'s) can be considered an "index of cohesiveness" among panelists' judgements, with greater cohesiveness evident for smaller numbers, and more "spread" in the data and generally less cohesiveness for larger numbers. Standard deviation results can only be directly compared between components for components having the same number of assessment indicators. The numbers in Table 2 have been rounded up or down accordingly to arrive at whole numbers for percentages and single decimal place numbers for standard deviations.

An analysis of the means and the standard deviations of the recommended percentages for components in each round of iterations revealed that, in almost all instances, standard deviations decreased from one round to the next round. In 18 of the 22 Teaching and Learning Components, the standard deviation decreased from the first round to the third and final round; in one component, the standard deviation stayed the same; in three components the standard deviation increased but only negligibly (.1). Further, in 14 of the 22 components, the standard deviation decreased progressively in each round of iterations.

This is consistent with the findings of Cross, Impara, Frary, and Jaeger (1984) that distributions become more homogeneous when panelists are given opportunities to reconsider their recommendations. Increased homogeneity of distributions increases the reliability of the recommended standards.

The greatest variance from the mean was seen for the six components making up Performance Dimension I, Preparation, Planning and Evaluation which is used in assessing the Comprehensive Unit Plan. Greater variability might possibly be explained by the fact that

panelists had had less training and experience in assessing this Dimension and therefore were less certain of performance expectations.

Considering the various recommendations collectively, mean percentages for the first round of recommendations typically approximated 75%. This finding shows that these panelists typically recommended that the percentage of "1's" in most STAR Teaching and Learning Component matrices should be approximately 75%. Notable exceptions to this typical standard were for component numbers: 10 (Managing Task-Related Behavior, 70%); 11 (Monitoring and Maintaining Student Behavior, 70%); 13 (Physical Learning Environment, 83%); 14 (Lesson and Activities Initiation, 71%); 18 (Thinking Skills, 67%); and 23 (Oral and Written Communication, 87%). The lowest recommended performance standard was for STAR component 18 (Thinking Skills, 67%) and the highest recommended standard was for STAR component 23 (Oral and Written Communication, 87%).

Table 3 presents the means and standard deviations of percentages of acceptable decisions recommended as performance standards by teachers in comparison to other position types on the standards-setting panel. The "other" category includes principals, parish supervisors, college faculty and state department of education personnel.

Given the "high stakes" teachers have as a group in the standards to be set, there was interest in whether teachers' recommendations would vary significantly from those of other professionals. As indicated by the results, recommendations were very consistent. On 13 of the 22 recommended performance standards, percentages were the same or within one point for the various groups; 20 of the 21 were within two points of each other. Only one standard differed by 3 percentage points, and that was for the Component of Thinking Skills, the component on

which pilot data showed the lowest level of STAR assessment scores (Claudet, Hill, Ellett, & Naik, 1990).

Standard deviations about the means suggest somewhat less cohesiveness in judgements among teachers than among other groups. Variability was greater for teachers on 17 of the 22 components. Differences were fairly small for most components, however, with the greatest spread shown for Component 13, Physical Learning Environment (S.D. = 6.08 for teachers; S.D. = 3.73 for others) and Component 4, Aids and Materials (S.D. =6.62 and 4.10 for teachers and others respectively).

As the results of this standard-setting task were being compiled and interpreted, the 1989-1990 version of the STAR was undergoing revision based on information from the two pilot years (1988-1990) of research and development in Louisiana. Therefore, the recommended standards shown in Table 2 had to be translated into the 1990-1991 revision and structure of the STAR that has currently been approved by the Louisiana Board of Elementary and Secondary Education (BESE). This revision in the STAR resulted in merging some assessment indicators with other components (a minor change) and deleting some assessment indicators that were considered somewhat "redundant." The total reduction in the number of assessment indicators was from 140 (1989-1990 version) to 117 (1990-1991 version). These revisions were made in an attempt to make the STAR assessment process more efficient. However, consideration was also given to the developing research base from studies of the STAR in Louisiana and to maintaining the professional integrity and quality of the STAR assessment process.

Table 4 shows the relationship between the structure of the 1989-1990 STAR and the 1990-1991 STAR revision and the application of the performance standards recommended by these

panelists. Included in this table are: 1) the number of assessment indicators comprising the various STAR teaching and Learning Components (#I's); 2) the size of assessment matrices for each component (number of assessment indicators X 3 assessors X 2 occasions; fall and spring); and 3) the percentage standard recommended along with the number of indicator decisions at this standard that should be assessed as "Acceptable." These data are shown for both the 1989-1990 and 1990-1991 versions of the STAR.

There are two issues reflected in the results in Table 4 when viewing recommended standards relative to the number of assessment indicators comprising components in the 1989-1990 and 1990-1991 versions of the STAR. First, with the reduction in the number of assessment indicators for some STAR Teaching and Learning Components, the number of allowable "0's" in a matrix has also been somewhat reduced. For example, Teaching and Learning Component #11 (Monitoring and Maintaining Student Behavior) has been reduced from 9 to 6 assessment indicators. This revision reduces the size of the assessment matrix (3 assessors X 9 assessment indicators X 2 occasions; fall and spring) for this component from 54 to 36 decisionmaking possibilities ("Acceptables" or "Unacceptables"). Applying the recommended standard of 70% to the matrices for component #11 shows that 16 assessment decisions could be "0's" in the 1989-1990 version of the STAR (i.e., 54 minus 38), while only 11 assessment decisions can be "0's" in the 1990-1991 version of the STAR (36 minus 25) if the current, recommended standard is to be met. Thus, the number of "1's" required to meet the recommended standard in the newer (1990-1991) version of the STAR (n=25) is less than what might have been required in the 1989-1990 version of the STAR (n=38). However, the newer (1990-1991) version of the STAR might be considered somewhat "tougher" because there is less margin for "Unacceptable" decisions in

the total component matrix than in the 1989-1990 version of the STAR.

Considering all of the Teaching and Learning Components for the two versions of the STAR, nine retained the same number of indicators and assessment matrix sizes. One component (Student Engagement) is not used to make certification decisions. Component #20 (Pace) was merged with other components in the 1990-1991 STAR revision. Twelve components were reduced in size in terms of the number of assessment indicators and the issues discussed above for component #11 (Monitoring and Maintaining Student Behavior) apply to each of these components. Thus, a reduction in the number of assessment indicators for the STAR Teaching and Learning Components from 1989-1990 to 1990-1991, when viewed relative to the performance standards recommended by this group of panelists, makes the STAR somewhat less "flexible" in terms of the "requirements" that must be met for certification. Further reductions in the number of assessment indicators in future versions of the STAR, given the recommended performance standards shown in Table 3 will create even more inflexibility in the STAR as a data collection and decisionmaking framework. If there are further reductions, "flexibility" in the STAR in making assessments for certification can only be accommodated by lowering performance standards for the Teaching and Learning Components. These issues are also important to consider in applying the STAR to the wide variety of classroom contexts (e.g, subject matter, class size, nature/characteristics of students, grade level, etc.) in which teaching and learning occur.

Secondly, with differences in the number of assessment indicators comprising the various components, and a common standard (e.g., 75%), there are fewer possibilities for allowable "0's" for components with a small number of assessment indicators than for those components with

a larger number of assessment indicators. For example, component #5 (Homework/Home Learning) was reduced in terms of the number of assessment indicators from 4 to 3. The number of possibilities for allowable "0's" (with a standard of 75%) dropped from 6 (24 minus 18) to 4 (18 minus 14). Whereas, component 16 (Aids and Materials) was reduced in terms of the number of assessment indicators from 8 to 6. The number of possibilities for allowable "0's" (with a standard of 75%) dropped from 12 (48 minus 36) to 9 (36 minus 27).

These observations about the current structure of the STAR as a decisionmaking framework might suggest a need to establish uniformity in terms of the number of "data points" in the decisionmaking matrices for the various Teaching and Learning Components. However, this requirement is not advisable since the assessment indicators have been professionally verified by Louisiana educators as "essential" elements of the components under which they are classified. The assessment indicators for the various components also have support in the existing research literature on effective teaching and learning. In addition, arbitrarily moving some assessment indicators from their current classification to other Teaching and Learning Components simply to achieve symmetry in instrument structure will create considerable difficulties in the conceptual basis of the STAR and in the professional development program to certify assessors. These changes would also affect the "face validity" of the STAR, the logical classification of the assessment indicators, and most importantly perhaps, the reliability of the various components for making certification decisions.

Given these concerns, it should be noted here, that regardless of the number of assessment indicators comprising a particular STAR Teaching and Learning Component, the recommended performance standards reflect the views of a group of Louisiana educators having full knowledge

of the effects of setting standards for components with differing numbers of indicators.

Insight into factors influencing recommendations of panelists was gained from a Participant "Importance to Recommendations' Scale completed by each panel member. Using a five point Liekert scale ranging from a 5 for "highly important" to a 1 for "little or no importance," panelists were asked to assess the degree to which eight items influenced their final recommendations about the appropriate performance standards for each STAR Teaching and Learning Component. Table 5 presents the frequency with which participants responded.

The importance of the STAR component for enhancing student's learning was identified as having had the greatest influence on the group as a whole with 85% ranking it highly important and an additional 13.3% ranking it above average in importance. Other items to which panelists attributed a high degree of importance were the consequences and impact on teachers of setting standards at various levels (66.7%, highly important; 26.7% above average in importance) and the training received as part of the STAR professional development program to certify assessors (57.8%, highly important; 28.9% above average in importance).

Feedback about the standards set by other participants and the discussion and rationale for standards provided by other participants appeared to have the least influence, but these factors were still viewed as above average or high in importance by 55.5% and 70.5% of respondents respectively.

Overall the impact on students and then teachers appeared to have the greatest influence on panelists' recommendation of performance standards followed by knowledge of results from STAR research data, and last by the views of other panelist members. Panelists apparently viewed some teaching and learning components as "more critical and important for certification,"

or as more "difficult" or "easy" given actual performance data from the 1989-1990 STAR pilot, than others...resulting in a reduction in the allowable "0's" for some Teaching and Learning Components relative to others. Therefore, recommended standards for some components were higher or lower than for other components.

As new findings from the literature on effective teaching and learning emerge, the STAR may need to be lengthened to accommodate these findings and to add additional flexibility to the system for decision making purposes and to assure applicability of the STAR to the wide variety of contexts in which teaching and learning occur.

One final activity for panelists was to consider final performance standards recommended for experienced teachers for professional, renewable certification in view of a "performance expectation" for new teachers participating in the Louisiana Teaching Internship Program. After reviewing final standards and entertaining considerable discussion, panelists recommended that a "voice vote" be taken to endorse the idea that the performance expectations for intern teachers, as a requirement for completing the internship program, should be the same as the recommended performance standards for professional, renewable certification for experienced teachers. The rationale for this recommendation was that a key purpose of the Internship Program should be to help prepare the new beginning teacher for new certification requirements, and that the intern teacher should not be considered as having met the requirements of the Internship Program without some assurances that certification standards for the renewable certificate would be met. Panelists were asked to write an attestation affirming their agreement with this voice vote. There was little argument with this recommendation, and all panelists agreed that the performance

expectation for intern teachers should be the same as the performance standard for experienced teachers for renewable certification.

Implications and Recommendations

This document has briefly described the results of an initial performance standards-setting task(s) with a select group of Louisiana educators for the Louisiana System for Teaching and learning Assessment and Review (STAR). The purpose of this task was to arrive at a set of recommendations for initial performance standards for experienced teachers seeking professional, renewable certification as required in the Children First Act (1988) and for beginning teachers to meet performance expectations as required in the Louisiana Teaching Internship Program Law (1984).

These recommendations were submitted to the Louisiana Department of Education (LDE) for review and for consideration by the Louisiana Board of Elementary and Secondary Education (BESE) along with a variety of issues and concerns that needed to be reviewed before a final set of performance standards were considered and adopted by the BESE. First, and foremost, it was suggested that the standards recommended by this panel of Louisiana educators should not be considered final by any means but be viewed as a set of "benchmarks" for teachers as they prepared for their fall, 1990 STAR assessments. A variety of factors supporting this approach were provided. Most of these factors had to do with the "assessment demand characteristics" under real conditions and possible changes in program implementation policies. As these factors change, predictable changes in STAR performance levels will be evident.

For example, it is known that the performance distributions of assessment systems like the STAR change under "high stakes" conditions like those reflected in assessments for professional,

renewable certification. These changes in performance can effect not only consideration of future standards levels, but the reliability of assessments and the "dependability" of the STAR for making certification decisions as well. In view of these predicted changes in performance, it was pointed out that the STAR performance data collected under pilot conditions did not involve preparation of the STAR Comprehensive Unit Plan (CUP). Research studies completed in the spring of 1990 suggest that teachers attempting and completing the CUP typically performed at higher levels in the actual classroom setting than teachers who did not prepare the CUP. This finding suggests that further improvements in performance levels in the fall of 1990 relative to performance levels evident during the 1989-1990 pilot year might be expected for those who prepare a CUP. Conversely, these findings may suggest that teachers who do not prepare the CUP as part of the STAR assessment may obtain assessment scores that are significantly lower than those who do.

The decision by the Board to require the CUP for all new beginning teachers in the Teaching Internship Program, and to not require the CUP for experienced teachers for renewable certification in light of the predicted effects on performance levels may raise "assessment equity" concerns and impact future standards-setting considerations for these two groups of teachers as well. Eventually the CUP, originally designed as an important, "reflective practice" part of the assessment process for certification, may not be scored at all because of lack of standardization across teachers assessed.

An additional, important policy-related concern that could drastically effect STAR performance distributions pertains to the "anonymity" of assessors. Research on teacher evaluation has clearly shown that assessment data are negatively skewed (inflated scores) if the

identity of assessors and their individual assessment decisions are known to teachers, or if the assessment team members are all "in-building" personnel. The social context of evaluation in this regard can put tremendous pressures on assessors to resist making "Unacceptable" decisions. The pilot process with the STAR was based on the policy of maintaining assessor anonymity and it was recommended that this policy be followed during actual program implementation in 1990-1991. However, tremendous pressure has been brought to change this policy.

This "assessor identity and score inflation" problem is also exacerbated if principals and other assessors must provide copies of raw observation notes to teachers. Pressure has been brought in this area as well. In the STAR pilot process, such notes were not part of the "evaluation,"... which is a summative decision relative to a standard...but, instead, they were simply used to guide the assessor in using the STAR Annotated Guide to make assessment decisions. The final "evaluation" is only made when all assessment data from all three assessors over two assessment occasions (fall and spring) are aggregated to make the final "evaluation" decision relative to an endorsed performance standard.

The mass of STAR data collected during the 1989-1990 statewide pilot (approximately 6,000 assessments) was collected under conditions of "everyday practice" and with the knowledge that results (assessment decisions and observation notes) did not have to be shared with teachers. These rules were followed because of the tentative nature of the pilot and because these field assessments were considered "practice" assessments for assessors. However, summaries of these data clearly show, that under the conditions of anonymity and everyday practice, the STAR, for the most part, clearly differentiates the quality of teaching and learning in Louisiana's classrooms (see 1989-1990 pilot results in Appendix B). Such differentiation among classrooms, and

"leaving room at the top," is also an important concern for identifying teachers as "superior" for the Model Career Options Program as required in the Children First Act. Protecting the anonymity of assessors seems an important issue from these perspectives.

The recommendations made by this group of panelists for teacher interns were made in view of the provision of adequate professional support from principals, master teachers and others. If these panelists had prior knowledge that there may be no support from master teachers within buildings (because current program logistics mitigate against this possibility), perhaps these panelists would have recommended a different set of standards for beginning teachers than for experienced teachers. With an "outside" master teacher instead of an in-building master teacher proposed as part of the assessment model for 1990-1991, recommendations may have been altered for experienced teachers as well. Belief by panelists in the kind of support that will be available for all teachers is an important factor in setting performance standards.

Thus, there are essentially two sets of important issues raised with these and other possible changes in program implementation policies as part of the STAR assessment: 1) changes expected in STAR performance distributions in the fall and spring of 1990-91 and their implications for future standards-setting considerations; and 2) the more global concern of policy changes such as whether careful, reflective practice should be part of a system in Louisiana to renew the professional teaching credential. The first concern, perhaps, awaits additional empirical information from assessments completed in 1990-91. The second concern is whether Louisiana's system will stay "in step" with future generations of assessment practices and methodologies being proposed and/or piloted by those working in the teacher certification field (e.g., the National Board of Professional Teaching Standards)...or whether policy changes will affect not

only standards but result in the replication of systems that have been developed and received "mixed blessings" in other states. The assessment of careful planning and reflective practice as an indication of content and pedagogical knowledge, for example, is currently viewed by the NBPTS and experts in the assessment field as an important assessment concern (NBPTS, 1990).

The standards recommended in this study were the "best" available based on the views of a panel of knowledgeable Louisiana educators consisting of a majority of classroom teachers. It was recommended that these standards be temporarily endorsed by BESE as appropriate "benchmarks" to guide teachers as they prepare for STAR assessments during the fall of 1990. It was also strongly recommended that standards be carefully reviewed and additional recommendations be made in the fall of 1990 and the spring of 1991 after data were analyzed that were collected under "real" conditions with final program implementation policies put into effect. Because of the tentative nature of the standards, it was further recommended that final decisions about professional, renewable certification not be made until the late spring of 1991, and that no teacher be considered "in remediation" until the fall of 1991. This procedure would allow those teachers who were initially randomly selected by the LDE for assessment to have a maximum of three years before any consideration could be given to invalidating the professional teaching credential. This seemed reasonable given the temporary nature of currently recommended standards reflected in this report and given the fact that some teachers would have much more time to prepare for the assessment process than other teachers.

Unfortunately, because teacher evaluation and standards-setting, must operate within a political and highly emotional arena, the most sound approach does not always take precedence. Responding to a great deal of pressure, the Louisiana Department of Education and the Board

of Elementary and Secondary Education made the decision to set final standards in December of 1990.

An external advisory committee of six consultants from within and outside the state met to review standards-setting issues, STAR research data and assist in establishing the final standards. After four days of meeting together the committee made a variety of recommendations relative to standards-setting concerns, decision-making models and elements of program implementation. A second set of three "external" consultants with extensive experience in the field of assessment reviewed these recommendations and suggested additional factors to be considered. The collective recommendations of both consultant groups were used in designing and implementing the Fall 1990 STAR standards-setting task.

Upon the recommendation of the external advisory groups, the standards-setting panel consisted entirely of classroom teachers, ten of whom had served on the original "expert panel" that set initial STAR benchmarks and the remaining group of 15 who had been assessed in the Fall 1990 STAR assessment. As part of a two-day workshop teachers considered various models and procedures for making assessment decisions, and teachers were able to review some of the fall data collected under high stakes conditions. These results showed "higher" scores under real versus pilot conditions. Higher scores have been interpreted to reflect actual improvement as well as artificial score inflation. (See Appendix D for a comparsion of pilot and fall data.)

Benchmarks established for each Teaching and Learning Component during the Spring of 1990 remained unchanged. However, teacher panelists recommended two substantive changes in the STAR decision-making model:

1. requiring that teachers achieve the benchmark on any 13 of the teaching components rather than all 15 as proposed in the original model, and

2. using a combined "conjunctive" and "compensatory" model such that the required numbers of Teaching and Learning Components must be at or exceed the "benchmark" standards and an overall assessment score of 75% should also be met.

The BESE accepted these recommendations, and standards were set accordingly. Since that time a number of ramifications have brought into question the decision to depart from the original criterion-referenced decision-making model. Not only do the changes diminish the construct validity of the STAR and the importance of each component in teaching and learning but it has also resulted in a number of "scoring anomalies," thus raising "equity" concerns. These concerns just reinforce the view that standards-setting is an ongoing process.

In future years, as LTIP and LTEP are implemented with all teachers, the LDE and BESE will need to revisit established performance standards for the STAR in view of: 1) performance data collected in Louisiana's classrooms under "real" conditions (a summary of the results of the completed evaluation of 1,701 teachers during the 1991-92 school year is included in Appendix D); 2) ramifications of policy decisions, ; 3) research results emerging from studies of the STAR and effective teaching and learning; and 4) the value that Louisiana educators and citizens place on particular elements of effective teaching, professional practice, and children's learning.

The following are just a few of the issues to be considered as standards are reviewed and revised:

How can we arrive at a set of standards for STAR that.....

(1) assure a "reasonable" level of competence at certification levels,

(2) will protect the "integrity" of new important skills such as content structure/emphasis and thinking skills,

(3) will drive staff development and professional improvement,

(4) will be useful for making Model Career Options decisions,

(5) will send and protect the "accountability" message desired by the Governor's Office, BESE, the legislature and the general public,

(6) will not create "supply and demand" problems in the teacher work force,

(7) will not "overload" principals with remediation work,

(8) will be clear and understandable to teachers, principals, policymakers and the general public,

(9) allow for "revisiting" and "adjusting" after a sufficient number of teachers have been assessed both Fall and Spring,

(10) will reflect what is empirically known about STAR validity and reliability characteristics,

(11) discourage score "solution," "corruption," and "inflation," realizing that no single model is likely to be able to completely control this "noise" in the system,

(12) will be useful in determining/examining "significant progress,"

(13) will be "non-discriminatory" against particular classes of teacher (e.g., race, gender, grade level, etc.),

(14) will reinforce the concepts of "criterion-referenced" assessment and "banking" for certification; and criterion-referenced assessment for MCOP (no banking),

(15) will allow the SDE with available monies to successfully manage the program (e.g., number of teachers in remediation ("carryovers"), number of available assessors, etc.)

(16) will be sensitive to the statistical "dependability" of decisions and will minimize "false negative" and "false positive" decisions (this relates to reliability concerns in #10 above), and

(17) will be conceptually consistent with the "holistic" construct validity of the STAR?

As stated in the introduction to this paper, setting performance standards for systems like the STAR is not an easy task. When considered in light of a volatile socio-political context, the task becomes even more complex.

References

Brandstatter, H., Davis, J.H. & Stocker-Kreichgauer, G. (1983). Group decision making. New York: Academic Press

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

Chauvin, S.W. & Ellett, C.D. (1990). Initial perceptions of educators in Louisiana regarding the Louisiana Teaching Internship and Statewide Teacher Evaluation Programs. Technical report number 3, Louisiana Teaching Internship and Teacher Evaluation Projects, College of Education, Louisiana State University, Baton Rouge, Louisiana.

Claudet, J.G., Hill, F.H., Ellett, C.D., and Naik, N.S. (July, 1990). Summary of statewide assessment data from the 1989-90 pilot of the System for Teaching and learning Assessment and Review (STAR). Technical Report number 2, Louisiana Teaching Internship and Teacher Evaluation Projects, College of Education, Louisiana State University, Baton Rouge, Louisiana

Ellett, C.D. (March, 1990). A 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.

Ellett, C.D. (March, 1990). Development and initial pilot test of the System for teaching and learning Assessment and Review (STAR) to meet requirements of the Louisiana Teaching Internship and Statewide Teacher Evaluation Programs: Final project report (FY 1988-1989). Louisiana Teaching Internship and Teacher Evaluation Project, College of Education, Louisiana State University, Baton Rouge, Louisiana.

Ellett, C.D., Chauvin, S.W., Loup, K.S. and Naik, N.S. (March, 1990). Summary of statewide assessment data from the 1989 pilot of the System for Teaching and Learning Assessment and Review (STAR). Technical report number 2, Louisiana Teaching Internship and Teacher Evaluation Projects, College of Education, Louisiana State University, Baton Rouge, Louisiana.

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

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

Ellett, C.D., Loup K.S., Chauvin, S.W. and Naik, N.S. (March, 1990). An initial investigation of the criterion-related validity of the System for Teaching and learning Assessment and Review (STAR). Technical report number 6, Louisiana Teaching Internship and Teacher Evaluation Projects, College Education, Louisiana State University, Baton Rouge, Louisiana.

Hambleton, R.K. (1987). On the use of cut-off scores with criterion-referenced tests in instructional settings. Journal of Educational Measurement, 15(4), 277-290.

Hambleton, R.K. & Eignor, D.R. (1980). Competency test development, validation, and standard setting. In R.M. Jaeger & C.K. Tittle (Eds.), Minimum competency achievement testing: Motives, models, measures, and consequences (pp. 367-396).

Jaeger, R.M. (1978). A proposal for setting a standard on the North Carolina High School Competency Test. Paper presented at the annual meeting of the North Carolina Association for Research in Education, Chapel Hill, North Carolina

Jaeger, R.M. (1982). An iterative structured judgment process for establishing standards on competency tests: Theory and application. Educational Evaluation and Policy Analysis, 4, 461-475.

Jaeger, R.M. (1989). Certification of student competence. In R.L. Linn (Ed.) Educational Measurement (3rd ed., pp. 485-514). New York: Macmillan.

Jaeger, R.M. (1990). Setting standards on teacher certification tests. In J. Millman and L.D. Hammond (Eds.) The new handbook of teacher evaluation: Assessing elementary and secondary school teachers. Newbury Park, California: Sage Publications, Inc.

LaMaster, S. Tobin, B. and Bowen, C. (June, 1990). Making sense of an innovation: A qualitative study of the pilot implementation of LTIP/LTEP in three schools. College of Education, Florida State University, Tallahassee, Florida.

Loup, K.S., Ellett, C.D. and Chauvin, S.W. (April, 1990). Confirmatory Analyses to establish the construct validity of an observational system to assess teaching and learning. Paper presented at the annual meeting of the American Educational Research Association, Boston, Massachusetts.

National Board for Professional Teaching Standards. Toward high and rigorous standards for the teaching profession: initial policies and perspectives of the National Board for Professional Teaching Standards; a summary. Washington D.C.: National Board for Professional Teaching Standards, 1990.

Shepard, L. (1980). Standard setting issues and methods. Applied Psychological Measurement, 4(4), 447-467.

Teddlie, C., Ellett, C.D., and Naik, N.S. (April, 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.

APPENDIX A

Data Tables

TABLE 1

LTIP AND LTEP INITIAL STANDARDS-SETTINGS COMMITTEE

PARTICIPANT DEMOGRAPHIC INFORMATION

Sex:	
Male	32.6%
Female	67.4%
Ethnicity:	
Black	39.1%
White	60.9%
Average Age:	42.2 Years
Current Position:	
Teachers	56.5%
Principals	6.5%
Assistant Principals	17.4%
College Faculty	
Other	
Average Number of Years Employed in Current Position	11.89 Years
School Level in Which Currently Working:	
Early Childhood	2.3%
Elementary	41.9%
Secondary/High School	18.6%
College	7.0%
Multiple School Levels	4.7%
Average Years Experience As An Educator:	
Public Schools	18.4 Years
Private Schools	0.6 Years
Teachers' Average Years Teaching Experience:	
Public Schools	17.28 Years
Private Schools	1.00 Years
Highest Degree Earned:	
Bachelor	0.0%
Master	13.0%
Master Plus Graduate Hours	30.4%
Plus 30 or Specialist	47.8%
Doctorate	8.7%

TABLE 2

Summary of Means and Standard Deviations of Percentages of "Acceptable" Decisions Recommended as a Performance Standard for Each STAR Teaching and Learning Component for Three Rounds of Panelists' Recommendation

STAR Teaching and Learning Components		Mean[a] and S.D.[b] by Rounds					
		1	1	2	2	3	3
1.	Goals and Objectives	68	12.5	73	8.6	75	6.3
2.	Teaching Methods and Learning Tasks	73	6.3	74	6.3	75	5.2
3.	Allocated time and Content Coverage	74	8.3	75	5.7	75	4.9
4.	Aids and Materials	75	8.4	77	7.2	76	5.5
5.	Homework	74	9.3	75	8.3	75	5.5
6.	Formal Assessment and Evaluation	74	8.5	72	8.6	74	5.6
7.	Time	77	5.9	74	6.9	75	4.1
8.	Classroom Routine	77	6.7	76	7.3	76	4.3
9.	Student Engagement[c]	--	--	--	--	--	--
10.	Managing Task-Related Behavior	67	7.2	68	7.1	70	6.3
11.	Monitoring and Maintaining Student Behavior	70	7.4	69	6.4	70	5.9
12.	Psychosocial Learning Environment	76	5.9	77	6.0	77	5.5
13.	Physical Learning Environment	86	5.1	85	5.2	83	5.2
14.	Lesson and Activities Initiation	67	7.9	69	6.8	71	4.8
15.	Teaching Methods and Learning Task	75	5.6	75	5.7	74	4.5
16.	Aids and Materials	76	5.0	76	5.4	75	5.1
17.	Content Accuracy and Emphasis	74	5.9	75	5.0	75	3.9
18.	Thinking Skills	65	10.5	65	9.7	67	8.2
19.	Clarification	75	4.2	76	4.0	75	4.3
20.	Pace	75	5.4	75	6.0	74	5.4
21.	Monitoring Learning Tasks and Informal Assessment	74	4.8	74	4.8	75	3.9
22.	Feedback	73	5.2	73	5.5	74	4.9
23.	Oral and Written Communications	87	6.3	88	6.0	87	4.9

[a]Mean = Arithmetical average computed over all panelists' performance standards recommendations

[b]S.D. = Standard Deviation of scores recommended by panelists

[c]Student Engagement standard was not recommended because this component is not used to make certification decisions

TABLE 3

Comparison of Means and Standard Deviations of Percentages of "Acceptable" Decisions Recommended as a Performance Standard for Each STAR Teaching and Learning Componet by Teacher Panelists and other Position Types

STAR Teaching and Learning Component		Teachers (n=26)		Others (n=18)	
		Mean[a]	S.D.[b]	Mean	S.D.
1.	Goals and Objectives	75	6.65	75	5.91
2.	Teaching Methods and Learning Tasks	75	5.49	75	5.10
3.	Allocated Time and Content Coverage	74	5.66	76	4.08
4.	Aids and Materials	75	6.62	76	4.10
5.	Homework	75	6.22	75	4.99
6.	Formal Assessment and Evaluation	74	6.41	74	4.05
7.	Time	75	3.29	76	5.20
8.	Classroom Routine	75	4.51	77	4.35
9.	Student Engagement[c]	--	--	--	--
10.	Managing Task-Related Behavior	69	5.78	71	6.95
11.	Monitoring and Maintaining Student Behavior	69	6.20	71	5.39
12.	Psychosocial Learning Environment	76	6.43	78	4.10
13.	Physical Learning Environment	82	6.60	83	3.58
14.	Lesson and Activities	71	4.48	72	5.16
15.	Teaching Methods and Learning Task	74	3.93	74	5.35
16.	Aids and Materials	75	5.59	76	5.02
17.	Content Accuracy and Emphasis	74	3.56	76	4.26
18.	Thinking Skills	65	8.26	68	8.15
19.	Clarification	74	4.36	76	4.04
20.	Pace	73	6.14	75	4.17
21.	Monintoring Learning Tasks and Informal Assessment	74	4.22	75	4.07
22.	Feedback	73	5.60	74	3.98
23.	Oral and Written Communications	86	6.08	87	3.73

[a]Mean = Arithmetical average computed over all panelists' performance standards recommendations

[b]S.D. = Standard Deviation of scores recommended by panelists

[c]Student Engagement standard was not recommended because this component is not used to make certification decisions

TABLE 4

Comparison of the Number of Assessment Indicators for 1989-1990 STAR and 1990-1991 STAR Teaching and Learning Components, Fall and Spring Assessment Matrix Sizes, and Recommended Performance Standard Percentages and Numbers

STAR Teaching and Learning Components			1989-1990			1990-1991		
			#I's[a]	MS	%(#)	#I's[a]	MS	%(#)
1.	Goals and Objectives	6	36	75(27)	4	24	75(18)	
2.	Teaching Methods and Learning Tasks	6	36	75(27)	4	24	75(18)	
3.	Allocated time and Content Coverage	4	24	75(18)	4	24	75(18)	
4.	Aids and Materials		5	30	76(23)	4	24	76(18)
5.	Homework ("Home Learning")[b]	4	24	75(18)	3	18	75(14)	
6.	Formal Assessment and Evaluation		7	42	74(31)	7	42	74(31)
7.	Time		8	48	75(36)	6	36	75(27)
8.	Classroom Routine		4	24	76(18)	4	24	76(18)
9.	Student Engagement[c]		-	-	-	-	-	-
10.	Managing Task-Related Behavior	6	36	70(25)	6	36	70(25)	
11.	Monitoring and Maintaining Student Behavior	9	54	70(38)	6	36	70(25)	
12.	Psychosocial Learning Environment	12	72	77(55)	10	60	77(46)	
13.	Physical Learning Environment	4	24	83(20)	3	18	83(15)	
14.	Lesson and Activities Initiation	10	60	71(43)	8	48	71(34)	
15.	Teaching Methods and Learning Task	6	36	74(27)	6	36	74(27)	
16.	Aids and Materials		8	48	75(36)	6	36	75(27)
17.	Content Accuracy and Emphasis	7	42	75(32)	6	36	75(27)	
18.	Thinking Skills	11	66	67(44)	11	66	67(44)	
19.	Clarification		5	30	75(33)	4	24	75(18)
20.	Pace[d]		3	18	74(13)	-	-	-
21.	Monitoring Learning Tasks and Informal Assessment		6	36	75(27)	6	36	75(27)
22.	Feedback		4	24	74(18)	4	24	74(18)
23.	Oral and Written Communications		4	24	87(21)	4	24	87(21)

[a]#I's = Number of assessment indicators comprising a STAR Teaching and Learning Component;

MS = Matrix size;

%(#) = Percentage standard recommended and number of assessment indicator decisions that must be "1's" (Acceptable)

TABLE 5

Frequency of the Degree to Which Items Influenced Final Recommendations of Performance Standards for the STAR Teaching and Learning Components for Experienced Teachers for Renewable Certification

	Items Influencing Recommendations	Degree of Importance[a] by Percentage of Respondents (n=45)				
		1	2	3	4	5
1.	The STAR professional development program to certify assessors.	0.0	8.9	4.4	28.9	57.8
2.	Direct experience in the field observing teachers with the STAR.	0.0	2.2	11.1	46.7	4.0
3.	Knowledge of results from STAR research data presented in this program.	0.0	2.2	22.2	37.8	37.8
4.	Knowledge of STAR reliability and decision-making concerns presented/discussed in this program.	0.0	2.2	17.8	42.2	37.8
5.	Feedback about standards set by other participants in this program.	2.2	13.3	28.9	22.2	33.3
6.	Discussion/rationale for standards provided by other participants.	0.0	15.9	13.6	43.2	27.3
7.	The consequences and impact on teachers of setting standards at various levels.	0.0	4.4	2.2	26.7	66.7
8.	The importance of the STAR component for enhancing students' learning.	0.0	0.0	2.2	13.3	84.4

[a]Importance scale

1 = Little or No Importance
2 = Some Importance
3 = Average Importance
4 = Above Average in Importance
5 = Highly Important

APPENDIX B

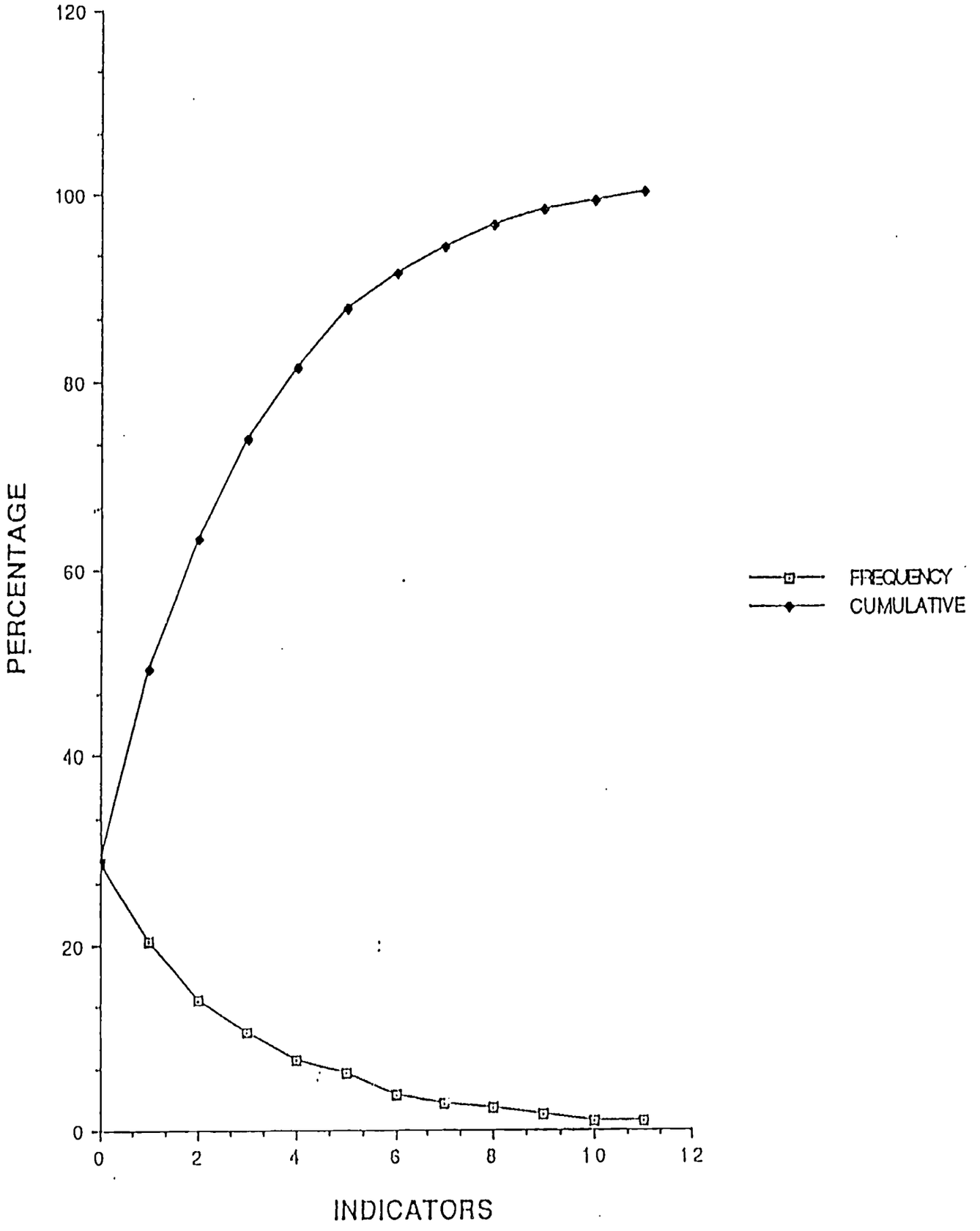
Sample Set of STAR Research Results Provided To Each Standards-Setting Panelist

TABLE 1
Percentage of Maximum Possible for Teaching and Learning Components for Each Dimension of the STAR
(Indicators = 108)
(N = 5720)

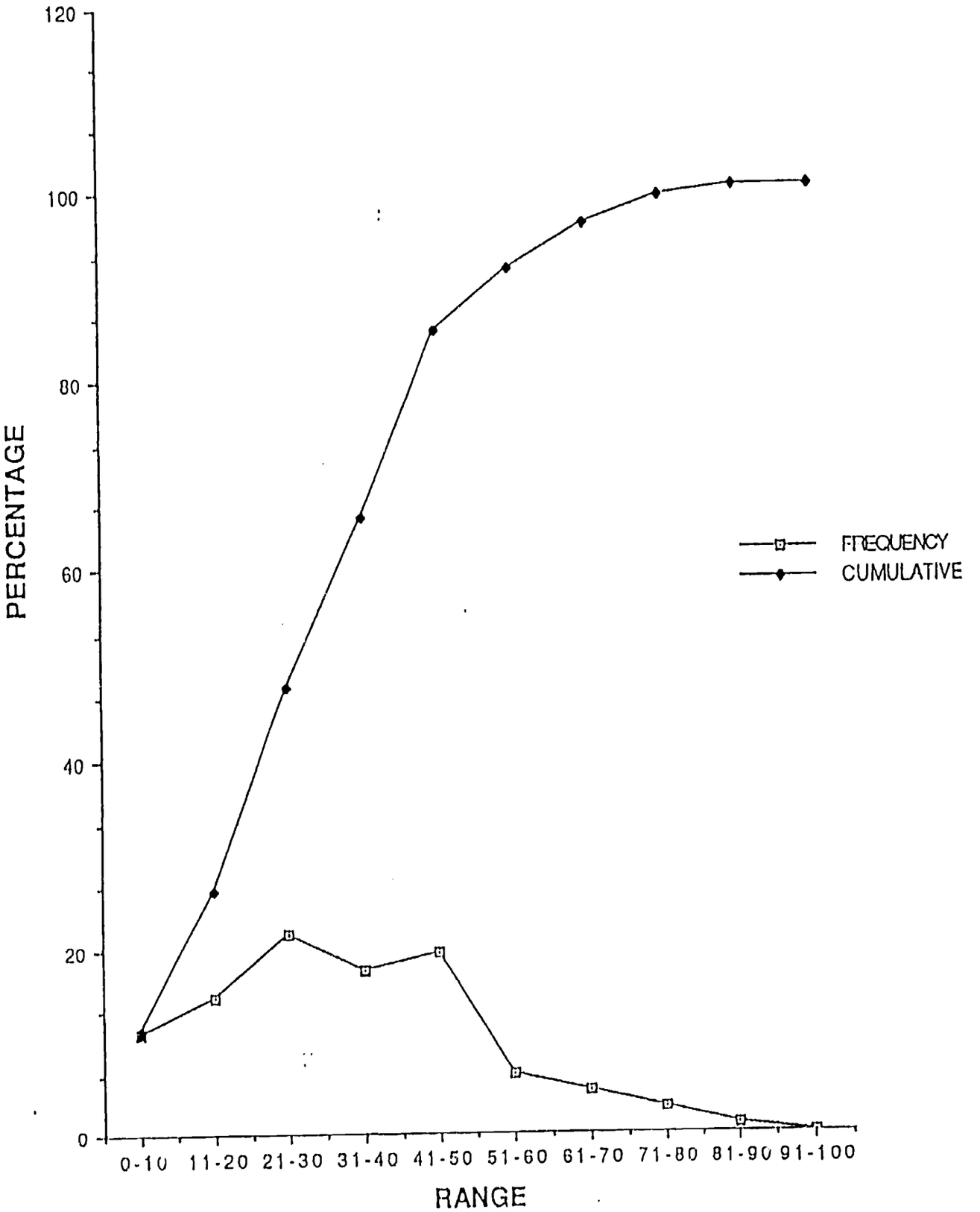
TEACHING AND LEARNING COMPONENTS		# of Indicators	Maximum Possible	% of Maximum
PERFORMANCE DIMENSION II: CLASSROOM AND BEHAVIOR MANAGEMENT				
A.	Time	8	43,784	72.39
B.	Classroom Routines	4	21,892	74.17
C.	Student Engagement	1	5,473	36.87
D.	Managing Task-Related Behavior	6	32,838	48.48
E.	Monitoring and Maintaining Student Behavior	9	49,257	54.21
PERFORMANCE DIMENSION III: LEARNING ENVIRONMENT				
A.	Psychosocial	12	65,676	66.40
B.	Physical	4	21,892	88.03
PERFORMANCE DIMENSION IV: ENHANCEMENT OF LEARNING				
A.	Lesson and Activities Initiation	10	54,730	34.45
B.	Teaching Methods and Learning Tasks	6	32,838	58.64
C.	Aids and Materials	8	43,784	61.78
D.	Content Accuracy and Emphasis	7	38,311	49.14
E.	Thinking Skills	11	60,203	21.56
F.	Clarification	5	27,365	54.28
G.	Pace	3	16,419	58.02
H.	Monitoring Learning Tasks and Informal Assessment	6	32,838	43.15
I.	Feedback	4	21,892	33.22
J.	Oral and Written Communication	4	21,892	94.70

TEACHING AND LEARNING COMPONENTS	Percent Acceptable			Percent Unacceptable		
	N	BT	ET	N	BT	ET
D. CONTENT ACCURACY AND EMPHASIS						
1. Students are given opportunities to learn at more than one cognitive and/or performance level.	32.0	31.3	32.3	88.0	68.7	67.7
2. Emphasizes the value and importance of topics and activities.	19.2	18.6	19.0	80.8	81.4	81.0
3. Content knowledge is accurate and up-to-date.	93.5	90.1	94.5	6.5	9.9	5.5
4. Content knowledge is logical.	87.2	82.8	88.1	12.8	17.2	11.9
5. Directions and explanations related to lesson content and/or learning tasks are effective.	57.1	53.1	57.3	42.9	46.9	42.7
6. Essential elements of content knowledge and/or performance tasks are emphasized.	28.7	25.3	28.9	71.3	74.7	71.1
7. Potential areas or points of difficulty are emphasized throughout the lesson.	26.2	22.6	26.3	73.8	77.4	73.7
E. THINKING SKILLS						
1. Associations are taught and used in learning.	26.3	26.1	26.6	73.7	73.9	73.4
2. Involves students in developing concepts.	26.1	25.0	26.7	73.9	75.0	73.3
3. Involves students in developing principles and/or rules.	15.3	14.5	15.3	84.7	85.5	84.7
4. Encourages students to think of and recall examples from their own experiences.	22.7	22.4	22.7	77.3	77.6	77.3
5. Encourages students to use mental imagery.	13.9	15.4	13.6	88.1	84.6	86.4
6. Asks a variety of questions.	29.3	31.3	28.8	70.7	68.7	71.2
7. Wait time is used to enhance student learning.	38.2	36.4	37.7	61.8	63.6	62.3
8. Encourages critical analysis and/or problem solving.	18.4	16.6	18.4	81.6	83.4	81.6
9. Encourages students to elaborate, extend or critique their own or other students' responses.	18.5	16.0	18.2	81.5	84.0	81.8
10. Encourages creative thinking.	14.8	15.4	14.2	85.2	84.6	85.8
11. Provides opportunities for the extension of learning to new contexts.	13.6	14.3	13.1	86.4	85.7	86.9

TEACHING & LEARNING COMPONENT
IV E: THINKING SKILLS



TEACHING & LEARNING COMPONENT
IV E: THINKING SKILLS



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

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	0.493
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

APPENDIX C

Hypothetical STAR Assessment Matrices for One Teacher for One Teaching and Learning Component

Hypothetical Data Matrices for a STAR Teaching and Learning Component for One Teacher for an Assessment Year for One Teaching and Learning Component Comprised of Eight Assessment Indicators

Fall Assessment

Indicators	P[a]	MT	Ext.
1	1	1	1
2	0	0	1
3	0	1	1
4	1	1	1
5	1	1	0
6	1	1	0
7	1	0	0
8	0	1	1
	5	6	5

Total "1's" = 16 (67%)

Spring Assessment

Indicator	P	MT	Ext.
1	0	1	1
2	1	0	1
3	1	1	0
4	1	1	1
5	1	0	1
6	1	1	1
7	1	1	0
8	0	0	0
	6	5	5

Total "1's" = 16 (67%)

GRAND PERCENTAGE = 67%[b]

[a]P = Principal; MT = Master Teacher; Ext = External Assessor

[b]Grand Percentage = Sum of "1's" for the two Matrices Divided by 48 Possibilities

APPENDIX D

Summary of STAR Data Analyses Evaluations Completed

Summary of Percentages of Teachers at or Exceeding "Benchmarks" for
STAR Teaching and Learning Components from Spring, 1990
"Pilot" (Single Assessor) Assessments and Fall, 1990 "Real" (All Three Assessors) Assessments

Teaching and Learning Components	Spring, 1990 (n = 5,720)	Fall, 1990 (n = 2,587)
1. Time (75) *	65	92.2
2. Routines (76)	69	95.3
3. Managing Task-Related Behavior (70)	33	89.0
4. Monitoring and Maintaining Student Behavior (70)	35	82.5
5. Psychosocial Learning Environment (77)	34	92.2
6. Physical Learning Environment (83)	68	94.9
7. Lesson and Activities Initiation (71)	7	58.3
8. Teaching Methods/Learning Tasks (74)	34	90.8
9. Aids and Materials (75)	49	90.1
10. Content Accuracy/Emphasis (75)	11	67.7
11. Thinking Skills (67)	6	38.4
12. Clarification (75)	40	91.0
13. Monitoring Learning Tasks and Informal Assessment (75)	20	84.8
14. Feedback (74)	22	73.1
15. Oral and Written Communication (87)	87	97.8

Recommended "Benchmark"

Summary of STAR Data Analysis Results for TEP Teachers 50
Completed on Existing SDE Data File as of February 22, 1991
(n=1,701 teachers)

1. "Satisfactory" Decisions

# Components Passed	# Teachers
13	151
14	377
15	991
	TOTAL = 1519

Percentage of teachers "passing" STAR with current requirements = 89.3%

* Note: The 1519 teachers are exactly the same teachers with or without the "compensation score" of 395. Therefore, the compensation score does little to contribute to the decision to certify.

2. "Superior" Decisions

# Components Passed	# Teachers
14	82
15	47
	TOTAL = 129

Percentage of teachers qualifying for MCOP relative to total number of teachers assessed in file = 7.6%

* Note: The number of teachers with a Master degree and at least 7 years teaching experience = 254. The 129 "qualifiers" for MCOP meeting standards is equal to 50.8% of the sample of 254 teachers.

3. The percentage of teachers "passing" all 15 components at the current benchmark standards is 58.3% (991/1701).

4. The percentage of teachers "passing" all 15 components at the current benchmark standards <u>with the standard for Thinking Skills lowered by 33 points</u> = 72.5%.

5. The percentages of teachers that "pass" either 13 or 14 components at current benchmarks and "fail" selected components are as follows:

"Pass" 13 Components

Thinking Skills	(122 of 151 "fail")	(80.8%)
Feedback	(28 of 151 "fail")	(18.5%)
Lesson Initiation	(65 of 151 "fail")	(43.0%)
Content Accuracy/ Emphasis	(45 of 151 "fail")	(29.8%)

"Pass" 14 Components

Thinking Skills	(267 of 377 "fail")	(70.8%)
Feedback	(23 of 377 "fail")	(6.1%)
Lesson Initiation	(36 of 377 "fail")	(9.5%)
Content Accuracy/ Emphasis	(21 of 377 "fail")	(5.6%).

* Note: A total of 389 teachers (73.7%) in these two groups of "certifiable" teachers "fail" to meet the current benchmarks for the STAR Thinking Skills Component.

* Note: The percentage of teachers "failing" to meet the current benchmark standard for the Thinking Skills component of all three groups of "certifiable" teachers (13,14 and 15 "passed") is 25.6%.

6. The percentage of teachers that "pass" either 11 or 12 components at current benchmarks and "fail" selected components are as follows:

"Pass" 12 Components

Thinking Skills	(63 of 69 "fail")	(91.3%)
Feedback	(24 of 69 "fail")	(34.8%)
Lesson Initiation	(49 of 69 "fail")	(71.0%)
Content Accuracy/	(39 of 69 "fail")	(56.5%)

* Note: If the "passing" model for certification is lowered to "any 12" components, the overall "pass" rate would increase from 89.3% to 93.4%.

"Pass" 11 Components[8]

Thinking Skills	(33 of 36 "fail")	(91.7%)
Feedback	(16 of 36 "fail")	(44.4%)
Lesson Initiation	(29 of 36 "fail")	(80.6%)
Content Accuracy/ Emphasis	(24 of 36 "fail")	(66.7%)

* Note: If the "passing" model is lowered to "any 11" components, the overall "pass" rate would increase from 93.4% (above example) to 95.5%.

* Note: If the "passing" model is lowered to "any 11" components, the overall "failure" rate on Thinking Skills for those passing fewer than 15 components would be 76.6%.

7. An analysis of a file of 7,787 complete fall, 1990 STAR assessments showed that 4,727 teachers (60.7%) were at or exceeded current benchmarks on 13 or more components. A similar analysis showed that 198 teachers (2.54%) were at or exceeded current benchmarks for MCOP on 14 or 15 components.

8. The following percentages are "failure" percentages for each STAR teaching and learning component based on analyses of the file of 1,701 complete assessments when compared to existing "benchmark" standards for each component:

STAR Component	% "Failures"
Time	1.6%
Routines	0.7%
Managing Task-Related Behavior	3.8%
Monitoring/Maintaining Behavior	6.4%
Psychosocial Learning Environment	1.2%
Physical Learning Environment	0.8%
Lesson Initiation	14.8%
Methods and Tasks	1.9%

Aids and Materials	2.1%
Content Accuracy/Emphasis	11.3%
Thinking Skills	32.8%
Clarification	3.2%
Monitoring LearningTasks/ Informal Assessment	4.2%
Feedback	8.7%
Oral and Written Communication	0.4%

9. The following percentages show "pass" rates for each STAR teaching and learning component for fall, 1990 and spring 1991 assessments for the sample of 1,701 teachers for each component.

STAR Component	Fall %	Spring%
Time	96.6	99.6
Routines	98.1	99.5
Managing Task-Related Behavior	92.8	98.2
Monitoring/Maintaining Behavior	90.1	96.7
Psychosocial Learning Environment	96.3	99.5
Physical Learning Environment	98.5	99.3
Lesson Initiation	70.6	92.8
Methods and Tasks	94.9	98.5
Aids and Materials	93.9	99.1
Content Accuracy/ Emphasis	79.5	96.9
Thinking Skills	45.7	88.4
Clarification	92.1	98.2

Monitoring Learning Tasks/ Informal Assessment	91.9	98.4
Feedback	85.0	97.9
Oral and Written Communication	99.4	99.6