Research in Standards-based Science Assessment

Kathleen B. Comfort, Principal Investigator
Mark Wilson, Co-Principal Investigator
WestEd, San Francisco, CA

Project Summary

RISSA, Research in Standards-based Science Assessment, proposes to conduct a research study to investigate what types of science assessment measures provide data on student outcomes that are useful, accurate, and allow valid inferences about student achievement. RISSA is a project of WestEd and the University of California, in collaboration with RAND, three organizations that share a reputation as successful leaders in educational reform.

The overall goal of the project is to help Systemic Initiatives, other NSF-supported science reform efforts, and others to (1) better measure the impact of their efforts to improve student learning and performance in science and (2) more effectively use assessment results to inform and improve instruction.

RISSA will investigate the following research questions: What type(s) of measures best inform standards-based instruction? Does exposure to inquiry-based science instruction improve scores on some components of standards-based science assessments more than others, and if so, which ones? How do teachers use data and results from assessment to inform practice? Will training in use of data change instruction? Is there differential reliability, validity and usefulness in reporting standards-based results by item type? Can we obtain useful information by equating a standards-based, criterion-referenced assessment with results of a norm-referenced assessment?

In order to examine these questions, RISSA will identify and recruit 96 schools (32 each, elementary, middle, and high school) 288 teachers (96 from each level); and 8640 students (2880 from each level). RISSA will use measures developed by the California Systemic Initiative Assessment Collaborative (CSIAC), an existing, NSF-supported science assessment. Data will be collected by means of pre- and post-tests as well as teacher and student interviews and surveys before the pre-test and after the post-test. Both student and teacher data will be analyzed using a multidimensional item response model which will then be used for the remainder of the analyses.

Results of the proposed study will (1) allow NSF-supported SI's and other science reform projects to make informed decisions about the usefulness of various types of measures for making sound inferences and promoting instructional change; and (2) to assist teachers in the use of assessment results to inform instruction and practice.

Project Description
1. Introduction

Purpose

As stated in the ROLE program announcement, many educational approaches, materials and technologies have been developed without the benefit of a strong research base. In many instances, this is because the appropriate research simply does not exist. While many studies have been conducted over the last ten years on the properties of science assessment, and on current approaches for the development and scoring of science assessments (Klein, Shavelson, Stecher, McCaffrey, Haertel, Comfort, Solano-Flores, & Jovanovic, 1996; Comfort, 1991, 1992, 1994, 1995, 1996; Baxter & Glaser, 1996; Wilson & Sloane, 2000; Stecher & Klein, 1995; Shavelson, Baxter & Gao, 1993; Klein, Stecher, Shavelson, McCaffrey, Ormseth, Bell, Comfort and Othman, 1998; Saner, Klein, Bell and Comfort, 1994), there has been limited research addressing the relationship(s) among types of science assessment, instruction, and student achievement.

The purpose of this three-year project, Research in Standards-based Assessment (RISSA), is to extend the current research base on science assessment by investigating timely and important issues within the framework of systemic initiatives (SI's) and other NSF-supported science reform efforts. Results of RISSA will help SI's and others determine what types of assessment provide data on student outcomes that are useful, accurate, and allow them to make valid inferences about student achievement. Valid and accurate assessment of student learning is a crucial component in judging the effects of science reform (NRC, 1996). Further, NSF expects SI's to accumulate a "broad and deep array of evidence that the program is enhancing student achievement through a set of indices." (NSF,1999). The proposed project will examine the efficacy of various measures in providing that evidence.

Results will help determine (1) what types of assessment best inform standards-based instruction; (2) whether and how teachers use assessment results to inform instruction; and (3) what reporting strategies are most useful and reliable for reporting standards-based results. Additionally, results will be used to examine how findings from this study can be used to inform the implementation and development of other large-scale assessment efforts. We will also attempt to address what kind of relationship might exist between the results of a criterion-referenced test, such as the California Systemic Initiative Assessment Collaborative (CSIAC)[1], and results of a norm-referenced standardized test such as the Stanford Achievement Test, version 9 (SAT-9). According to a National Research Council (NRC) report on standards-based assessment and accountability, assessments should involve a range of strategies appropriate for valid inferences relevant to individual students, classrooms, districts, and states (Elmore & Rothman, 1999). An important result of this proposed study will be to: (1) allow NSF-supported SI's and other science reform projects to make informed decisions about the usefulness of various types of measures[2] for making sound inferences and promoting instructional change; and (2) to assist teachers in the use of assessment results to inform instruction and practice.

Goals

[1] An NSF-supported standards-based science assessment (award no. nsf-9816805)
[2] measures are defined as different types of assessment components.

WestEd and the University of California at Berkeley, in collaboration with RAND, propose to form RISSA to engage in a process of assessment research. The project's overall goal is to help systemic initiative programs (SI's) and other NSF-supported science reform projects to (1) better measure the impact of their efforts to improve student learning and performance in science and (2) more effectively use assessment results to inform and improve instruction. Outcomes will be widely disseminated at academic, administrative, and practitioner levels by means of journal and newsletter articles and presentations at national, regional, and local conferences.

RISSA proposes to use an experimental instrument based on an existing standards-based science assessment that includes four different types of assessment components (CSIAC, described in a later section). In order to ensure that our research design is coherent and complete, as part of our proposed research activities we will design, pilot test, and administer two additional item types. The experimental version will then be used in RISSA's proposed research in order to more closely align to the national standards and to determine the efficacy of various types of measures and mixtures of measures within the context of SI's and other science reform efforts. The specific research questions to be addressed as well as project design and methodology are described in the following sections.

The three collaborating organizations share a reputation as successful leaders in educational reform, and have worked productively on other NSF projects. Through this unique collaboration and in keeping with the expectations described in the ROLE program announcement, the activities and outcomes of this project will (1) provide information useful to ongoing efforts to improve the practice of science education in the classroom; (2) establish strong connections between research and practice that will help to advance systemic reform efforts nationally; and (3) expand the base of research in the area of large-scale assessment. RISSA will collaborate with reform leaders in SI's and in other NSF-supported science reform projects to conduct its research. Results will directly impact classroom practice both in terms of assessment choices and instructional change.

RISSA is responding to Quadrant III., Research on SMET Learning in Educational Settings, with connections to Quadrant IV., Research on SMET Learning in Complex Educational Systems.

2. Statement of Need

The release of the National Research Council's National Science Education Standards (NSES) in 1996 as well as the American Association for the Advancement of Science Project 2061's Benchmarks for Science Literacy (BSL) in 1993 represented the national consensus of the science education community about what was important and meaningful for all students to know, do and understand in science.

In response, many states, districts and schools quickly initiated efforts to reform the quality of their science programs. They expected these efforts to lead to improved learning, which in turn would lead to high achievement in science by all students. However, they were faced with a serious problem—how would they know if their students were learning the content recommended by national science standards?

As the primary feedback mechanism in the educational system, assessment communicates the goals that students, teachers, schools and districts are expected to achieve (NRC, 1996). Research shows that students learn what they are taught and that teachers teach what they are

held accountable for (Shavelson, Carey & Webb, 1990; Resnick & Resnick, 1992; and Walker & Schaffarzick, 1974). In order to achieve educational reform, then, assessments must be aligned to the content recommendations of the national standards. Alignment ensures that tests match the learning goals embodied in the standards, and thus enables the public to determine student progress toward the standards (NRC, 1999; Linn & Herman, 1997; Webb, 1997; Wiggins, 1989; Raizen & Kaser, 1989; Anderson, 1990; and Hein, 1990).

Many states, districts and schools, especially those involved in NSF systemic initiative programs and other NSF-supported science reform projects, expressed a need for a valid and reliable standards-based assessment that would allow them to: measure student progress toward science literacy against the national standards; provide meaningful and useful data and results as well as strategies for using the results; and inform curriculum, instruction and teaching practices. In order to assist these projects with the development, administration, scoring and reporting of valid and reliable standards-based science assessments at the elementary, middle and secondary levels, NSF funded the California Systemic Initiatives Assessment Collaborative (CSIAC) in 1996. The work of CSIAC builds on current research on the properties of science assessment and current approaches for assessment development and scoring (Klein, Shavelson, Stecher, McCaffrey, Haertel, Comfort, Solano-Flores, & Jovanovic, 1996; Comfort, 1991, 1992, 1994, 1995, 1996; Stecher & Klein, 1995; Shavelson, Baxter & Gao, 1993). All CSIAC assessments are: aligned to the content recommendations of the NSES and the BSL; consist of a balance of measures—enhanced multiple-choice questions, open-ended questions, constructed response investigations, and hands-on performance tasks—that tap into different aspects of knowledge (Baxter & Shavelson, 1994); and are available in English and Spanish. To date, the CSIAC assessments have been administered to over 170,000 students in 16 states and Puerto Rico. In addition to providing standards-based science assessments, CSIAC has also provided an ongoing, organized forum for SI's and other NSF-supported science reform efforts to discuss and address common assessment issues and educational practices. These discussions suggest that research on valid and reliable standards-based assessments is a vital need among programs in the NSF systemic initiative portfolio.

The proposed RISSA project will utilize the work of CSIAC in order to extend research technologies and expand the existing knowledge base about student learning and assessment practices by investigating questions centered on: (1) the relationship among assessment, instructional practices and student achievement; (2) teacher understanding and use of assessment results to inform instruction and change classroom practice; (3) the reporting of standards-based results; and (4) how the research addressed in this investigation might inform other large-scale science assessment efforts.

Little is known about which type of assessment components best assess and inform standards-based science instruction, i.e., which measures provide most useful and robust information to educators as they continue their efforts to reform science education and improve science learning. Sites engaged in science reform efforts often ask what types of measures—such as multiple-choice items, open-ended questions, constructed response investigations, or hands-on performance tasks—are most useful for informing standards-based instruction, e.g., instruction that is inquiry-based. Another common question involves hands-on performance tasks: are recipe-type tasks or experimental tasks in which students design their own investigation more sensitive to instruction (Stecher, et al., 2000)?

A prominent rationale for using performance assessments to measure student performance is the belief that "you get what you assess... [and] you do not get what you do not assess" (Resnick and Resnick, 1992). This phrase acknowledges that performance assessments serve both a measurement function and signaling function. Advocates hope that performance assessments will encourage changes in teachers' beliefs about assessment and their instructional practices (Flexer & Gerstner, 1993). For example, performance assessments might lead teachers to employ new curriculum content (e.g., complete tasks rather than disconnected bits of information) and new instructional strategies (e.g., learning in context rather than decontextualized knowledge). Moreover, it is hoped that performance assessments will avoid the problems associated with high-stakes multiple-choice tests, including score-inflation and curriculum-narrowing (Koretz, et al., 1996; Shepard & Dougherty, 1991; Smith and Rottenberg, 1991). There has been some evidence that alternative assessments can make a difference (e.g., Wilson and Sloane (2000) report substantial gains compared to ordinary assessment practices), but there has been no systematic research on the ways that different types of assessment may be useful to teachers in designing and carrying out instruction.

Further, there is much speculation and some evidence that multiple-choice tests often fail to capture important abilities that may be assessed through more open-ended formats (Frederikson, 1984; Shavelson, Carey, & Webb, 1990), but much of the discussion surrounding the limitations of multiple-choice tests fails to acknowledge their strengths (Hambleton & Murphy, 1992; Mehrens, 1992; Stiggins, 1994). A well-constructed multiple-choice test may be a valuable component of an assessment system because it can provide broad coverage of important topics and allow students to demonstrate a variety of skills and knowledge. Moreover, it may be that the optimum choice will consist of mixtures of item types, as argued by Wilson and Adams (1996), and exemplified in Wilson and Wang (1996).

We will investigate these questions about the usefulness of different types of measures and different mixes of measures through the research activities outlined in this proposal. We will begin to address the question of whether or not there is an optimum mix of assessment modes within a context of standards-based, inquiry-based instruction by looking at different modes to see if a relative differential exists in terms of both student achievement and usability of assessment data by teachers.

A second area of concern is whether, how and why teachers use results from assessment to actually change their instructional strategies and content in educationally meaningful ways—not simply "teaching to the test." Teachers are central to implementing the vision of the national standards and to keeping curriculum, instruction and assessment closely linked. Current research maintains that teacher involvement in assessment development and scoring stimulates teachers thinking about their curricular vision and about how different instructional approaches can support students' learning (Shingold, Heller & Paulukonis, 1995; Shepard, 1995). When teachers are provided the opportunity to discuss and learn how to use assessment data in conjunction with their instruction, they have greater knowledge of what their students can do (Shepard, 1995), and their behavior differs from those of teachers without such information (Roberts, 1996). Through the design, implementation and analysis of survey results, interviews and observations, we are proposing to investigate the degree to which teachers use assessment results to inform instruction and change practice if they are provided the opportunity to participate in action research and training linked to their classrooms.

Norm-referenced tests (NRT's), developed by test publishers, measure student performance against a norming sample. Describing what students can do relative to other students, NRT's are used to compare groups of students. Criterion-referenced tests (CRT's), on the other hand, compare student performance to a set of established criteria (e.g., national standards) rather than comparing them to the performance of other students (Bracey, 1998). While reform sites want to know how well students are achieving the standards, they are also under increasing pressure to provide data allowing national and local comparison of student performance. RISSA intends to investigate the possibility of equating the results of our test administration with results from the same students taking the SAT-9 in California. Additionally, RISSA will also investigate other standards-based reporting strategies, e.g., the reliability and usability of reporting results via individual assessment modes as opposed to reporting assessment results via one score for a combination of different modes.

Finally, RISSA's research on the combination of different item types, scaling and equating studies, reporting strategies, and development of methods for using data to inform instruction will inform efforts to improve and enhance other large-scale assessments such as TIMSS and NAEP. For example, CSIAC has already done preliminary work by assisting the Merck Local Systemic Change Center, a four-district reform consortium, with revising, administering, scoring and reporting results for a TIMSS release task. Results from the task were reported at the classroom, school, district and overall Merck levels, and teachers were trained in how to use the data to inform their instruction.

3. Research Questions

RISSA intends to investigate the following research questions:

I. Issues regarding the relationship among assessment, instruction and student achievement
- What type(s) of measures—enhanced multiple-choice questions [EMC], open-ended questions [OEQ], constructed response investigations [CRI], hands-on performance tasks [HPT], experimental design problems [EDP], complex multiple-choice questions [CMCQ], or mixed measures [MM]—best inform standards-based instruction?
- Does exposure to inquiry-based science instruction improve scores on some components of standards-based science assessments more than others? Which ones? (E.g., primarily performance components?]

II. Issues regarding teacher understanding and use of assessment results to inform instruction
- How do teachers use data and results from assessment to inform practice: Does their instruction change as a result of learning how to use data and participating in RISSA training in use of data? Do experienced and inexperienced teachers respond differently to training in data use?

III. Issues related to the reporting of standards-based results
- Is there differential reliability, validity and usefulness in reporting standards-based results by item type?
- Can we obtain useful information by equating results of a standards-based, criterion-referenced assessment (experimental version of CSIAC) with results of a norm-referenced assessment (SAT9)?

IV. Extension of assessment related issues to other large-scale assessments
- How will this research inform other large-scale science assessments, such as TIMSS? Here we will address questions relating to: combination of item types; scaling and equating; reporting of results; and use of data to inform instruction.

4. Program Methods

In order to investigate the questions listed above, RISSA will include the following research components: (1) identification and recruitment of participants and treatments; (2) development and pilot testing of instrumentation and assessment components; (3) implementation and administration of instrumentation and assessment components; (4) training teachers in data use; (5) data collection; (6) procedures and data analysis; and (7) dissemination.

Identification and Recruitment of Participants and Treatments

The RISSA sample population will include 288 teachers (96 elementary, 96 middle school, and 96 high school); 96 schools (32 elementary, 32 middle school and 32 high school); and 8,640 students (2880 elementary [5th grade], 2880 middle school [8th grade], and 2880 high school [10th grade]). All of these teachers, by virtue of their involvement in CSIAC, will be participating in inquiry-based instruction, but there will be considerable variation in the quality of that instruction (which we will investigate through teacher surveys and interviews).

RISSA will identify and recruit teachers from systemic initiative programs and other NSF-supported science reform efforts in California. Although this research study will be conducted in California, results will be generalizable to SI's and other science reform efforts throughout the country. Collaborating projects will include the San Francisco Unified School District (an Urban Systemic Program), the Oakland Unified School District (a Comprehensive Partnership for Mathematics and Science Achievement, Lasers, (a Local Systemic Change Center), and the California K-12 Alliance, which includes SPAN (NSF-supported project).

Participating teachers will be requested to: complete and return all survey instruments; participate in interviews and observations; administer pre- and post- assessments to one class of students; and participate in on-site discussions with colleagues administering similar assessment components. Teachers will also participate in the training on use of data to inform instruction.

The pre-assessment (Form A) will be administered to a representative sample of 8,640 students across three grade levels (5th, 8th and 10th); as will the post-assessment (Form B). Both forms will consist of six types of items: (1) enhanced multiple-choice questions, (2) open-ended questions, (3) constructed response investigations, (4) hands-on performance tasks, (5) complex multiple-choice questions, and (6) experimental design problems. The first four types have been developed, pilot and field tested, and widely implemented as part of CSIAC (described later in this section). The complex multiple-choice questions and experimental design problems will be developed as part of this research project in order to be included as part of the design for the proposed treatments.

Between the pre-assessment and the post-assessment, teachers will be trained in the use of information from assessments, and given feedback from their students' performance according to

one of the conditions. The conditions are: one for each of the six assessment types listed above, and two more, corresponding to two different levels of mixed assessments: one will consist of all six item types, and one will consist of three item types. We will decide on which three after pilot-testing. Four schools, each with three teachers and approximately 90 students per school, will be assigned to each assessment condition. Across the entire set of science teachers at each grade level, we will also seek to recruit representative proportions of new and experienced teachers. The following chart shows the assignment of assessment types to schools for one grade level; it will be identical for the other two grade levels.

Assessment Type	Number of Schools	Number of Teachers/School	Number of Students/School	Total Schools	Total Teachers	Total Students
Enhanced Multiple-choice Questions	4	3	90	4	12	360
Open-ended Questions	4	3	90	4	12	360
Constructed Response Investigations	4	3	90	4	12	360
Performance Tasks	4	3	90	4	12	360
Complex Multiple-choice Questions	4	3	90	4	12	360
Experimental Design Problems	4	3	90	4	12	360
Mixed Assessments-1	4	3	90	4	12	360
Mixed Assessments-2	4	3	90	4	12	360
Totals:				32	96	2,880

Development of Instrumentation and Research Assessment Components

During the first year of the project, RISSA will develop and pilot test: research measures for the two additional item types, and extra items of the other four types if they are needed to supply two linked forms; survey instruments and interview and observation protocols for teachers (pre- and post-assessment); and an opportunity-to-learn survey for students. The teacher surveys and interview and observation instruments will be designed to gather a variety of information about teachers' backgrounds and their current curriculum and instructional program. They will also be used to look at the effects of assessment on teachers' beliefs about teaching and learning and on changes in classroom practice as a result of the availability of assessment information. Questions on the surveys will focus on the content of the science curriculum, the nature of science instruction, and beliefs about the role of assessment. Questions will address: the types of curriculum/instruction teachers are implementing; how long they have been implementing it; how much training they have received in it; how it aligns to NSES/BSL; if it is inquiry-based; how they are assessing it and how they think it should be assessed; what is their experience with assessment—have they participated in development; what assessment instrument do they use; what kind of measures they think would best assess student understanding of standards-based

science and why; and whether they use data to inform instruction and how they do it. RAND staff have studied the effects of assessment reform in Vermont, Kentucky and Washington, which will provide a head-start for survey design (Stecher et. al., 1998). These surveys will be administered early in the second year of the project, prior to the administration of pre-assessment activities to students. The student opportunity-to-learn survey will contain questions similar to those implemented by NAEP and TIMSS, focusing on the type of science learned; how students are learning it; time spent on science homework, and so on. The student opportunity-to-learn survey will be administered to students as part of the pre-assessment activities.

RISSA will use and expand upon measures developed by the California Systemic Initiatives Assessment Collaborative (CSIAC) for the pre- and post- assessments. Since its funding by NSF in 1996, CSIAC has assisted systemic initiative programs and other science reform efforts with standards-based science assessment. Over the last four years, CSIAC has developed an extensive item bank consisting of the following assessment components at the 5th, 8th and 10th grade levels:

Assessment Component	Description	
Enhanced Multiple-choice Items (EMC)	A matrix approach samples across students and content standards. Six forms per grade level.	Assess students' understanding of important scientific facts, concepts, principles, laws and theories.
Hands-on Performance Tasks (PT)	Investigations identifying a problem to solve. Answers are recorded in a test booklet and scored with a rubric.	Students use equipment to: perform investigations; make observations; generate, record and analyze data.
Open-ended Questions (OEQ)	Students are presented with a problem. They construct their own responses that are scored with a rubric.	Explore students' abilities to communicate scientific information and use science to express positions on societal issues.
Constructed Response Investigations (CRI)	Similar to performance tasks, but does not require equipment. Answers are recorded in a test booklet and scored with a rubric.	Students analyze a problem, conduct a secondary analysis, revise a hypothesis, and/or recommend solutions.
Student Opportunity-to-Learn Survey (SOLS)	Located on the back of the Student Information Form. Machine scorable.	NAEP-like questions addressing: the type of science learned; how they are learning it; time spent on science homework; and use of technology.
Student Information Form (SIF)	One-page machine-read document.	To collect: student information; EMC answers; and responses to SOLS.

All CSIAC assessment components are aligned to the content recommendations of the NSES and to the BSL. Good technical data exists for all components, and scaling and equating studies were completed to produce scale scores McCaffrey, Hamilton, & Aronson, 1998; Wilson, Delgado & Finklestein, 1998) so that participating states, districts and schools could examine growth and performance from year-to-year.

While the SI's use CSIAC as an instrument to measure their growth against the standards and to report program impact, RISSA is proposing to use an experimental version of the CSIAC

assessment to investigate the following questions: (1) what type of measures best inform and align with standards-based instruction and (2) which of these measures or mix of measures are more sensitive to assessing scientific inquiry.

In order to include a comprehensive range of measures in our proposed research, RISSA will create an experimental version of the CSIAC assessment by augmenting the existing four types of measures--enhanced multiple-choice items, open-ended questions, constructed response investigations, and hands-on performance tasks--with two additional research assessment components: complex multiple-choice questions (CMCQ) and experimental design problems (EDP). Regular multiple-choice items usually focus on assessing small, topical pieces of information such as, what are the parts of a plant, or in what year was helium discovered. Enhanced multiple-choice items are aligned to national standards and are designed to focus more on incorporating problem-solving skills with concepts and bigger ideas of science. Complex multiple-choice questions (CMCQ) will be designed to include a set of questions that will explore standards and concepts at a deeper level. For example, a student may respond to a CMCQ by selecting one of four predetermined answers. After selecting an answer, the student will be directed along a pathway of other items designed to probe deeper into student understanding or misconceptions. At the conclusion of the pathway, students will be requested to justify and explain why they selected their answers (Tamir, 1993).

The experimental design problems (EDP) will be similar to hands-on performance tasks. In most hands-on performance tasks students are presented with a problem to solve and a standardized set of directions instructing them how to solve the problem. For example, students are requested to gather specific types of data; record this data in a pre-designed table; draw conclusions based on the data; and in some cases, to make an application beyond the task. The EDP's however, will be designed using a more open format. Instead of being provided with recipelike directions, students will be presented with a problem and asked to design and carry out their own investigations in order to solve it (Stecher, et al., 2000).

A set of item shells will be used for the development of the CMCQ's and the EDP's. As described by Solano-Flores, et al. (1999), "Item shells are hollow frameworks whose syntactic structures generate sets of similar items (Klein, Stecher, McCaffrey, & Haertel, 1997b; Haladyna & Shindoll, 1989) or templates that specify the characteristics of families or types of problems (Hively, et al., 1968). By specifying the structural properties and formal characteristics of the CMCQ's and the EDP's, shells will help to ensure the comparability of these measures from year to year. Additionally, shells will help to initiate systematic discussions among the developers regarding the characteristics; skills and knowledge that the assessments intend to address (Solano-Flores, Jovanovic, Shavelson, & Bachman, 1999). Both CMCQ and EDP's will be developed interactively, with a try-out-review-revise approach (Solano-Flores & Shavelson, 1997). The items and tasks, response formats and scoring systems will be refined in each iteration. The assessments will be pilot- and field-tested to evaluate their content, psychometric soundness and usability. Scoring rubrics will be developed and pilot tested along with the EDP's.

In addition to the methods described above, we will conduct student interviews and think-aloud protocols with students as they work through the different assessment components (Hamilton, 1998). The student discussions will help to shed light on important issues and provide a record of the cognitive demand of each assessment component (Hamilton, Nussbaum & Snow, 1997; Magone, Cai, Silver & Wang, 1994).

Grade-level teachers from SI's and other NSF-supported science reform efforts in California will form teacher research teams to assist RISSA measurement specialists, researchers and scientists in the development and pilot testing of the CMCQ's and EDP's. All research-designed assessments will be developed and piloted tested in the first year of the project in a small sample of classrooms. Members of the RISSA team will observe teacher administration of the assessments and student participation in the pilot tests. Additionally, RISSA team members will interview students regarding their understanding of the new measures using think-aloud protocols.

Implementation of teacher surveys and administration of pre-assessment research components

All teachers participating in RISSA will be asked to complete the survey instrument at the beginning of the second year of the project, prior to the administration of the pre-assessments. During the fall of the second year of the project, 8,640 students (2880 at grade 5, 2880 at grade 8, and 2880 at grade 10) in 96 schools at each grade level will be administered the full battery of the experimental version of the CSIAC assessments (Form A). The experimental version, (XCSIAC) will consist of six assessment components—enhanced multiple-choice questions, open-ended questions, constructed response investigations, performance tasks, complex multiple-choice questions and experimental design problems.

All teachers across all sites administering the research-designed assessments will receive standardized training in test set-up and administration. A test window will be scheduled for administration. During the administration of the pre-assessments, a sample of teachers will be observed, and students will be interviewed using the think-aloud protocols. Following completion of the assessments, all testing materials will be collected. The RISSA research team will train and calibrate a group of readers to score the performance components. The multiple-choice sections of the test will be scanned for correct answers and scores from the performance components will be data entered. Results will be linked for individual students, classrooms, schools and districts. All data will be analyzed and preliminary reports of results will be produced. Reports of results will be used to train teachers in the use of assessment data.

Training in data use

Following the pre-assessments, scoring and data analysis during the fall of year 2 of the project, teachers will be assigned and trained in one of the eight assessment treatments. All teachers in the sample will be given general descriptive information about the results of the pre-assessment. After the initial surveys and the pre-test, one-half of the teachers will be trained in practical methods regarding (1) how to interpret information from the different assessment types, and (2) how to apply that new information to develop and/or change instructional strategies and techniques. After the post-test, this sample of teachers will be interviewed to determine what they actually did in the classroom and how useful they found the training. This training will take place during the winter of the second year of the project.

Administration of post-assessment research components and implementation of teacher interviews

Following the teacher's training in the different assessment components, students will be post-tested using Form B of XCSIAC. All teachers across all sites will receive refresher training in test set-up and administration. A test window will be scheduled for administration. During the administration of the post-assessments, a sample of teachers will be observed, and students will be interviewed using the think-aloud protocols. Following completion of the post-assessments, all

testing materials will be collected. The RISSA research team will use the same group of readers who scored the Form A performance components. The multiple-choice sections of the test will be scanned for correct answers and scores from the performance components will be data entered. Results will be linked for individual students, classrooms, schools and districts. All data will be analyzed and preliminary reports of results will be produced.

Data collection and scoring

Student responses to the pre- and post- assessments will be collected and returned to the RISSA testing contractor. This will include information on students' responses to the SAT-9 collected as part of participant district testing. Multiple-choice answers will be scanned for correct answers, and student responses to performance measures will be scored by teachers trained and calibrated in the scoring criteria. Scoring will follow procedures developed and implemented by CSIAC over the past four years; performance measures will be scored by readers trained and calibrated by the RISSA team.

All data, linked via bar code labels for students, classrooms, schools, districts, states and projects, will be entered, and data will be prepared for analysis. Students will be identified by numbers rather than by names. RISSA will ensure that the data collection methods do not invade the privacy of students, teachers, or parents. Additionally, RISSA will comply with state, district and/or school regulations for written permission from parents prior to testing, if permission is necessary.

Procedures and Data Analysis

The design described above will result in a two-level data set. Level one is comprised of the student-level data (pre- and post-test on the XCSIAC instrument, including the six assessment mode components, and student information from the pre-survey and opportunity-to-learn survey); level two is teacher-level data (information from the pre- and post surveys of teachers). In addition, teachers are clustered into schools, and schools are each assigned to one of the treatment/non-treatment conditions.

Following the pre-testing, dimensional analysis of the XCSIAC components will be carried out using a multidimensional item response model (Adams, Wilson & Wang, 1997; Wu, Adams & Wilson, 1998). Potentially, from this data, it could be determined that each of the six item mode components, and the composite of all six, each form a separate, identifiable and useable dimension (Wilson & Wang 1995). At the other extreme, it could be determined that, although the components constitute different information, that information falls along a common dimension. More likely, the result will be somewhere in between these extremes. The determination will be made by (a) significance testing, and (b) effect size estimation, on the hierarchical dimensional models. This process will result in a specific dimensional item response model that will be used for the remainder of the analyses. It will also be possible to make dimensional analyses of these with respect to the NRT information.

Common items on the two forms of the XCSIAC instrument (Forms A and B) will be used to link pre-test results to post-test results. Each of the eight component conditions (i.e., the six components plus the two mixed components) will be used as a condition in several multilevel analyses. Reliability indices will be calculated for the six components and the full composite, as well as interesting sub-composites. Several sorts of validity information will be available, including (a) comparison to teacher judgement (i.e., teachers will be asked to provide judgements of student

abilities as part of the teacher survey), (b) comparison with a NRT (SAT-9); and (c) information gathered from student interviews and think-alouds. In addition, the teacher post- survey and interview will give information on the usefulness of assessment information to instruction.

The ConQuest software (Wu, Adams & Wilson, 1998) allows one to incorporate these hypotheses directly into the analysis in the form of dummy-variable effects, where the no-treatment condition acts as the reference category. There are several interesting hypotheses that will need to be investigated. First, one can ask whether each of the six treatment-components make a detectable difference in terms of their specific test-components. One can ask the same question in terms of the composite variable; the mixed treatment-components will be included here as well. One can also ask a similar set of questions about detectable differences in the other test-components due to each treatment component (including the mixed components). The sorts of specific questions that will be addressed by this sequence of analyses will include: Do the multiple-choice items register an effect from the training teachers received in interpreting and using multiple choice items? Does the XCSIAC composite register an effect from the training that teachers received on open-ended items? Do the constructed response investigations register an effect from any of the other treatments? Having established where there are interesting differences, one can then proceed to ask (using a multilevel analysis) exploratory questions about whether teacher background and teacher instruction variables are related to any of these detectable differences. Teacher background variables will be collected in the teacher survey, and will consist of characteristics like years of experience and familiarity with assessments. Teacher instructional variables will be constructed during the course of the study, based on the variety of data we will collect from them. Although the exact nature of these is a matter for investigation during the course of the project, we can speculate that relevant variables might be something like: "Level of teacher usage of assessment results" (i.e., no feedback to students; scores or grades; scores with interpretations; scores with suggestions for next steps).

Dissemination

RISSA intends to engage in a comprehensive dissemination effort in keeping with the substantial need for the kind of results the proposed study will produce. We will submit articles for publication in the journals such as the Journal of Educational Measurement, Applied Psychological Measurement, Educational Evaluation and Policy Analysis, Educational Leadership, and journals published by NARST and NSTA We will develop and conduct presentations at local, regional, and national conferences for academic leaders, educational administrators, and science education practitioners, including the American Education Research Association (AERA), the Council of Chief State School Officers (CCSSO), the National Association for Research in Science Teaching (NARST), the National Science Teachers Association (NSTA), and other professional educational/measurement/research oriented conferences. We will conduct workshops for teachers and teacher leaders through professional organizations, systemic initiatives, and other science reform efforts. Information about the project and its results as well as planned presentations and workshops will be posted on WestEd's web site and on EdGateway, WestEd's unique interactive internet portal for the science and mathematics education community.

Activities Timeline

Year 1	Year 2	Year 3
• Refine and finalize designs,	• Train/instruct teachers in	• Collect data.

treatments, etc. Advisory Board input at beginning of project. • Identify, recruit and schedule participants and schools. Assign treatments. Schedule Year 1 & 2 activities. • Design, develop and pilot test instrumentation, protocols and research-based assessment components. • Score and analyze results from pilot tests. • Refine and finalize instrumentation and research-based assessment components. • Design and pilot test instruction and administration manuals for instrumentation and research-based assessment components. • Develop, pilot test and finalize training protocols for the 8 treatments. • Make arrangements for gathering SAT-9 information. • Train and calibrate readers in scoring criteria.	surveys. • Administer surveys to teachers. • Collect surveys and analyze data—we need to know about curriculum, etc, before we assign treatments? • Train teachers in test administration and set-up. • Administer experimental CSIAC pre-assessments to students. • Collect materials from schools—scan MC and CMC; prepare performance measures for scoring. • Conduct scoring. • Analyze data and produce preliminary reports. • Administer experimental CSIAC post-assessment to students—one of the eight treatments. • Collect materials from schools and SAT-9 data. • Train and calibrate readers in scoring criteria. • Score. • Data entry and analysis. • Produce reports. • Interview teachers and collect data.	• Analyze data. • Write reports. • Dissemination activities.

Advisory Board

The RISSA Advisory Board—similar in expertise to the one that advised CSIAC—will consist of researchers, measurement specialists, scientists, SI leaders, teachers and representatives from NSF-supported science reform projects. RISSA will also seek recommendations from NSF.

Project Management and Personnel

RISSA will be housed at WestEd's headquarters in San Francisco. Kathy Comfort, WestEd, will serve as Principal Investigator and Project Director. Mark Wilson, University of California, Berkeley, and the Berkeley Evaluation and Assessment Research Center (BEAR), will serve as Co-principal Investigator, and Tamara Kushner, WestEd, will serve as Senior Research Associate.

As co-PI/PD, Kathy Comfort will direct and coordinate RISSA activities and will collaborate with Mark Wilson and BEAR associates at UC Berkeley in: the design, development, pilot testing and administration of all instrumentation, protocols and research-based assessment

components; design and implementation of training for scoring and data use; development and implementation of research designs and treatments; site visits to collect data—interviewing teachers and students and making field observations; designing strategies for data collection and reporting; producing reports of findings; communicating with NSF; and disseminating research results at professional meetings, conferences, and teacher institutes.

In addition to substantial participation in the design and implementation of the previously mentioned activities, Mark Wilson will also design and conduct all data analysis, including the scaling and equating of NRT's with the standards-based instruments.

Tamara Kushner will provide professional support for all activities mentioned above and will also direct and coordinate all field work including the identification and recruitment of research sites, teachers and students; pre- and post assessment administration; teacher interviews and observations. She will also be responsible for coordinating all dissemination activities and all communication with partner schools and the Advisory Board. She will also assist with the writing of results and findings. An administrative assistant, housed at WestEd, will provide technical support for all activities listed above.

Stephen Klein, Laura Hamilton and Stephen Klein from RAND will assist with the design of instrumentation and research-based assessment components and with data analysis issues.

Results from Prior NSF Support

The proposed RISSA project will use the California Systemic Initiatives Assessment Collaborative (CSIAC)[3] as its primary student data collection instrument. CSIAC is described earlier in this section. It was funded by NSF from 1996 through 1999 as a development project to assist NSF-funded Systemic Initiative projects with the development, administration, scoring and reporting of valid and reliable standards-based science assessments at the elementary, middle and secondary levels. At the school, district and site levels, science reform projects using CSIAC have reported using results to: gather base-line and trend data; monitor progress and growth over time; and to inform participants of how well students are achieving the content recommended by the NSES and Benchmarks. Some sites have also reported using CSIAC results for accountability purposes where CSIAC data are allocated points, and reported in a school's report card.

The work of CSIAC was accomplished by an Advisory Board in conjunction with three Development Teams. The teams consisted of grade-level teachers and science specialists from participating SI programs. The Advisory Board, consisting of participating SI leaders and directors, scientists, measurement specialists, testing contractors, business and industry representatives, and university representatives, oversaw the progress of the Development Teams, ensured the scientific and psychometric quality of all items and tasks, and monitored the progress of the administration, scoring, and reporting of the assessments.

CSIAC engaged in work with RAND and the University of California, Berkeley, to design and conduct various technical studies. These studies included: equating and scaling test forms; initiating studies of reader consistency and score reliability; and establishing correlations among measures.

[3] Award number ESR-9632273, $999,895.00, November 1996-September 1998; and extended through September 1999, award number ESR-9816805, $307,308.00 .

The CSIAC assessments are valid in that they measure the content recommended by the *National Science Education Standards* and the *Benchmarks for Science Literacy*. CSIAC users from partner systemic initiative programs participate in the development of the assessments. They are in strong consensus on the match of the CSAIC assessments to national standards, and on the match of the CSIAC assessments to their local standards. Additionally, results from the CSAIC assessments are reported via the *National Science Education Standards*.

CSIAC has worked constantly to improve the reliability of all measures as well as the design of the test. In analyzing test results, the testing contractor produced item analyses, item bias analyses, statistical summaries and scale scores. Data from these analyses were used in the construction of all final forms, questions and tasks.

Common methods of item analysis were used to review the performance of all items on the multiple-choice tests, including p-values, point-biserial correlation coefficients and standard errors of measurement for all forms. Item response theory models are also used to identify items that perform poorly and to estimate reliabilities at different levels of ability. Differential item functioning using the Mantel-Haenszel procedure is used to investigate item functioning for different subgroups of the population.

For the performance measures (open-ended questions, constructed response investigations, and hands-on performance tasks), analytical scoring rubrics were developed and piloted with the initial versions of the questions and tasks. All readers were trained and calibrated to the criteria in the rubrics prior to scoring student work. The intra- and inter- reliability of reader consistency for the CSIAC assessments is considered to be high--.9. In addition to training and calibration, CSIAC implements consensus sets and read-behinds during scoring sessions to check for consistency. Data from both procedures show a high percentage of agreement and high reliability.

CSIAC has been used in four large-scale administrations (1997, 1998, 1999, and 2000) with over 167,000 students in nineteen states and Puerto Rico. Participants included: Urban Systemic Initiatives, 47%; Local Systemic Change Centers, 24%; Statewide Systemic Initiatives, 22%; Rural Systemic Initiatives, 2%; Comprehensive Partnerships for Mathematics and Science, 3%; and other NSF supported projects, 2%. Approximately 83% of students tested were from underrepresented groups and 17% were white, non-Hispanic.

References

Adams, R.J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement,* 21(1), 1-23.

American Association for the Advancement of Science. (1993). *Benchmarks for Science Literacy*. New York: Oxford University Press. Inc.

Anderson, R (1990) *California: The State of Assessment* (Sacramento: California Department of Education.

Baxter, G.P. and Shavelson, R.J., (1994). Science performance assessments: Benchmarks and surrogates. *International Journal of Educational Research,* 21, 279-298.

Bernauer, J.A. and Cress, K. (1997). How School Communities Can Help Redefine Accountability Assessment. *Phi Delta Kappan,* 79 (1), 71-75.

Comfort, K. B. (1996). Student Outcomes in Science, in N. Webb, ed., *Evaluation Strategies Working Group - National Institute for Science Education*. Wisconsin Center for Education Research, Madison, WI.

Comfort, K.B. and Michelson, P. (1995). *A Sampler of Science Assessment*. Sacramento: California Department of Education.

Comfort, K.B. (1994). Authentic Assessment: A Systemic Approach in California. *Science and Children,* 32(2), 42-43, 65-66.

Comfort, K.B., McCarthy, J., Trulson, M., and Michelson, P. (1994). *A Sampler of Elementary Science Assessment,* California Learning Assessment System, Sacramento: California Department of Education.

Comfort, K.B. (1992). New Directions in Science Assessment in California, in J. Duggins, ed., *Review* . San Francisco: San Francisco State University School of Education.

Comfort, K.B. (1992). Crime Solvers in California, in C. Lal and S.Rakow, eds., *Science Scope.* Washington, D.C.: National Science Teachers Association.

Comfort, K.B. (1991). A National Standing Ovation for the New Performance Testing, in G. Kulm and S.M. Malcom, eds., *Science Assessment in the Service of Reform,* Washington, D.C.: American Association for the Advancement of Science.

Flexer, R. J. and Gerstner, E. A. (1993, October). Dilemmas and issues for teacher developing performance assessments in mathematics: *A case study of the effects of alternative assessment in instruction, student learning and accountability practices* (CSE Technical Report 364). Los Angeles: Center for Research on Evaluation, Standards, and Student Testing.

Frederiksen, J.R., & Collins, A. (1989). A Systems Approach to Educational Testing. *Educational Researcher,* 18(9), 27-32.

Haladyna, T.M. and Shindoll, R.R. (1989). Shells: a method for writing effective multiple-choice test items. *Evaluation and the Health Professions,* 12, 97-104.

Hambleton, R. K. and Murphy, E. (1992). A psychometric perspective on authentic measurement. *Applied Measurement in Education, 5,* 1-16.

Hamilton, L.S. (1998). Gender differences on high school science achievement tests: Do format and content matter? *Educational Evaluation and Policy Analysis, 20,* 179-195.

Hamilton, L. S., Nussbaum, E. M., and Snow, R. E. (1997). Interview procedures for validating science assessments. *Applied Measurement in Education, 10, 181-200.*

Hein, G. (1990). *The Assessment of Hands-on Elementary Science Programs.* (Grand Forks, North Dakota: Center for Teaching and Learning. University of North Dakota.

Hively, W., Patterson, H.L., and Page, S.H. (1968). A 'universe-defined' system of arithmetic achievement tests. *Journal of Educational Measurement,* 5(4), 275-290.

Linn, R.L., and Herman, J.L (1997). *A Policymaker's Guide to Standards-led Assessment.* Denver, CO: National Center for Research on Evaluation, Standards and Student Testting.

Klein, S., Hamilton, L., McCaffrey, D., Stecher, B., Robyn, A., & Burroughs, D. (in press). *Teaching practices and student achievement: report of first-year results from the Mosaic study of systemic initiatives in mathematics and science.* RAND: Santa Monica, CA.

Klein, S., Stecher, B., McCaffrey, D., Jovanovic, J., Shavelson, R.,Haertel, E., Solano Flores, G., and Comfort, K. (1997) Gender and Racial/Ethnic Differences on Performance Assessments in Science. *Educational Evaluation and Policy Analysis,* 19(2), 83-97.

Klein, S., Shavelson, R., Stecher, B., McCaffrey, D., Haertel, E., Comfort, K., Solano-Flores, W., and Jovanovic, J. (1996). Sources of task sampling variability in science performance assessment tasks. Manuscript in progress.

Klein, S., Stecher, B., Shavelson, R., McCaffrey, D., Ormseth, T., Bell, R., Comfort, K., and Othman, A.R. (1996). Analytic versus holistic scoring of science performance tasks. *Applied Measurement in Education,* 11(2), 121-137.

Klein, S. P., Shavelson, R. J., Stecher, B. M., McCaffrey, D., and Haertel, E. (1997b). *Shell effects on hands-on tasks.* Manuscript submitted for publication.

Koretz, D.M., Barron, S. Mitchell, K. J., and Stecher, B.M. (1996). *Perceived effects of the Kentucky Instructional Results Information System* (KIRIS). Santa Monica: RAND Institute on Education and Training.

Magone, M., Cai, J., Silver, E. A., and Wang, N. (1994). Validating the cognitive complexity and content quality of a mathematics performance assessment. *International Journal of Educational Research, 21,* 317-340.

Mehrens, W. A. (1992). Using performance assessment for accountability purposes. *Educational Measurement: Issues and Practice, 11* (1), 3-9, 20.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-104). New York: Macmillan.

McCaffrey, D., Hamilton, L., and Aronson, E. (1998) *Scaling and Equating of the California Systemic Initiatives Assessment Collaborative Standards-based Assessments for 1996 and 1997:* Unpublished Technical Report, RAND: Santa Monica, CA

National Research Council. (1999). *Testing, Teaching and Learning.* Washington. DC: National Academy Press. (Committee on Title I Testing and Assessment. Elmore, R. and Rothman, R., eds.)

National Research Council. (1996). *National Science Education Standards.* Washington. DC: National Academy Press.

National Science Foundation, WWW.her.nsf.gov/her/esr/driver, November 1999.

Raizen, S., and Kaser, J., (1989). Assessing Science Learning in Elementary School: Why? What? How? *Phi Delta Kappan,* 70, 718.

Resnick, L. B. and Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In Gifford, B. and O'Connor, M. C. (Eds.), *Future assessments: Changing views of aptitude, achievement, and instruction*. Boston: Kluwer.

Roberts, L.L.C. (1996) Methods of evaluation for Public Understanding Program. Unpublished doctoral dissertation, University of California at Berkeley.

Saner, H., Klein, S., Bell, R., and Comfort, K. B. (1994) The Utility of Multiple Raters and Tasks in Science Performance Assessments. *Educational Assessment,* 2(3), 257-272.

Sheingold, K., Heller, J., and Paulukonis, S. *Actively Seeking Evidence: Teacher Change through Assessment Development.* (Princeton, NJ: Center for Performance Assessment, Educational Testing Services, 1995.)

Shavelson, R. J., Baxter, G. P., and Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement.* 30, 215-232.

Shavelson, R. J., Carey, N. B., and Webb, N. M. (1990). Indicators of science achievement: Options for a powerful policy instrument. *Phi Delta Kappan, 71*, 692-697.

Shepard, L.A.(1995) Using Assessment to Improve Learning. Educational Leadership, 52(5) 38-43.

Shepard, L. A. and Dougherty, K. C. (1991). *Effects of high-stakes testing on instruction.* Presentation at the annual meeting of the American Educational Research Association, Chicago.

Smith, M. L. and Rottenberg, C. (1991). Unintended consequences of external testing in elementary schools. *Educational Measurement: Issues and Practice,* 10 (4), 7-11.

Solano-Flores, G., Jovanovic, J., Shavelson, R.J., and Bachman, M. (1999). On the development and evaluation of a shell for generating science performance assessments. *International Journal of Science Education.* In press.

Solano-Flores, G. and Shavelson, R.J. (1997). Development of performance assessments in science: conceptual, practical and logistical issues. (Eng.) *Educational Measurement: Issues and Practice,* 16 (3), 16-25.

Stecher, B., Klein, S., Solano-Flores, G., McCaffrey, D., Robyn, A., Shavelson, R., & Haertal, E. (2000). The effects of content format and inquiry level on science performace assessment scores. *Applied Measurement in Education,* V 13 (2), p. 139-160.

Stecher, B.M. and Klein, S.P. (Eds.). (1995). *Performance assessments in science: Hands-on tasks and scoring guides.* Santa Monica, CA:RAND.

Stecher, B. M. and Mitchell, K. J. (1995). *Portfolio driven reform; Vermont teachers' understanding of mathematical problem solving* (CSE Technical Report 400). Los Angeles: Center for Research on Evaluation, Standards, and Student Testing.

Stecher, B. M. and Herman, J. L. (1997). Using portfolios for large scale assessment. In G. D. Phye (Ed.) *Handbook of classroom assessment* (pp. 491-517). San Diego: Academic Press.

Tamir, Pinchas (1993) Positive and Multiple-choice Items: How Different are They? *Studies in Education Evaluation* 19(3) 311-325.

Walker, D.F., and Schaffarzick, J. (1974). Comparing Curricula. *Review of Educational Research.* 44(1), 83-111.

Webb, N. (1997). *Determining Alignment of Expectations and Assessments in Mathematics and Science Education.* Madison, WI: National Center for Improving Science Education.

Wilson, M. & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education,* 13(2), 181-208.

Wilson, M., Delgado, J., and Finkelstein, D. (1998) *Equating and Scaling of the CSIAC Standards-based Science Assessments 1997 and 1998.* Unpublished Technical Report, Educational Data Systems: Morgan Hill, CA.

Wilson, M., & Adams, R.J. (1996). Evaluating progress with alternative assessments: A model for Chapter 1. In M.B. Kane (Ed.), *Implementing performance assessment: Promise, problems and challenges.* Hillsdale, NJ: Erlbaum.

Wilson, M., & Wang, W. (1995). Complex composites: Issues that arise in combining different modes of assessment. *Applied Psychological Measurement.* 19(1), 51-72.

Wiggins, G., (1989). A True Test: Toward More Authentic and Equitable Assessment: *Phi Delta Kappan*, 70, 703.

Wu, M., Adams, R.J., & Wilson, M. (1998). ConQuest [Computer Program]. Hawthorn, Australia: ACER.