Instructional Tools in Educational Measurement and Statistics (ITEMS) for School Personnel: Evaluation of Three Web-Based Training Modules

Rebecca Zwick, *University of California, Santa Barbara,*
Jeffrey C. Sklar, *California State Polytechnic University, San Luis Obispo,*
Graham Wakefield, *University of California, Santa Barbara,*
Cris Hamilton, *Independent Consultant,*
Alex Norman, *University of California, Santa Barbara,* and
Douglas Folsom, *University of California, Santa Barbara*

In the current No Child Left Behind era, K-12 teachers and principals are expected to have a sophisticated understanding of standardized test results, use them to improve instruction, and communicate them to others. The goal of our project, funded by the National Science Foundation, was to develop and evaluate three Web-based instructional modules in educational measurement and statistics to help school personnel acquire the "assessment literacy" required for these roles. Our first module, "What's the Score?" was administered in 2005 to 113 educators who also completed an assessment literacy quiz. Viewing the module had a small but statistically significant positive effect on quiz scores. Our second module, "What Test Scores Do and Don't Tell Us," administered in 2006 to 104 educators, was even more effective, primarily among teacher education students. In evaluating our third module, "What's the Difference?" we were able to recruit only 33 participants. Although those who saw the module before taking the quiz outperformed those who did not, results were not statistically significant. Now that the research phase is complete, all ITEMS instructional materials are freely available on our Website.

Keywords: assessment literacy, educational measurement, educational statistics, No Child Left Behind, psychometrics, standardized tests

According to a 2002 report from the Center on Education Policy, "[s]chool leaders . . . must be adept at using data to improve teaching and learning. Too often, test data is used only for accountability, rather than to diagnose the needs of individual students and improve their education. Both principals and teachers must be trained to use test data to modify daily lesson plans and to tailor assistance for individual children. Much more test data will soon be available, because the new federal requirements require states to produce descriptive and diagnostic information, as well as individual test results" (Jennings, 2002). As the report predicted, teachers and school administrators in the current No Child Left Behind (NCLB) era are expected to have a sophisticated understanding of test results, to use them to make data-based decisions about classroom instruction, and to communicate them to others.

Although there is little formal research in this area, it is widely recognized that many school personnel have not had the opportunity to acquire the "assessment literacy" that is required for these roles. Little has

Rebecca Zwick is a Professor, Department of Education, University of California, Santa Barbara, Santa Barbara, CA 93106; rzwick@education.ucsb.edu. Jeffrey C. Sklar is an Instructor, Department of CSM-Statistics, California State Polytechnic University, San Luis Obispo, CA 93407. Graham Wakefield is a PhD Candidate, Department of Media Arts and Technology, University of California, Santa Barbara. Cris Hamilton is an independent consultant. At the time of writing, Alex Norman and Douglas Folsom were both graduate students in the Department of Media Arts and Technology, University of California, Santa Barbara.

changed, in fact since the publication in 1990 of the *Standards for Teacher Competence in Educational Assessment of Students* (American Federation of Teachers, National Council on Measurement in Education [NCME], & National Education Association, 1990), which expressed concern about "the inadequacy with which teachers are prepared for assessing the educational progress of their students." The document recommended that "assessment training . . . be widely available to practicing teachers through staff development programs at the district and building levels" (p. 1).[1]

Further evidence regarding the need for training in this area comes from two additional efforts undertaken by professional organizations during the 1990s. The Joint Committee on Competency Standards in Student Assessment for Educational Administrators, sponsored by the American Association of School Administrators, National Association of Elementary School Principals, National Association of Secondary School Principals, and NCME, surveyed 1,700 administrators from the first three of these sponsoring organizations about assessment-related skills. The three skills rated as most needed by educational administrators (out of a list of 37) were knowing the terminology associated with standardized tests, knowing the purposes of different kinds of testing, and understanding the linkage between curriculum content and various kinds of tests (Impara, 1993, p. 20). In 1995, NCME published the *Code of Professional Responsibilities in Educational Measurement*. This document, which is intended to apply to professionals involved in all aspects of educational assessment, including those who administer assessments and use assessment results, includes the responsibility to "maintain and improve . . . professional competence in educational assessment" (NCME, 1995, p. 1).

According to Stiggins (2002), however, "only a few states explicitly require competence in assessment as a condition for being licensed to teach. No licensing examination now in place at the state or federal level verifies competence in assessment. Since teacher-preparation programs are designed to prepare candidates for certification under these terms, the vast majority of programs fail to provide the assessment literacy required to prepare teachers to face emerging classroom-assessment challenges . . . Furthermore, lest we believe that teachers can turn to their principals for help, almost no states require competence in assessment for licensure as a principal or school administrator at any level. As a result, assessment training is almost nonexistent in administrator-training programs."

The continuing need for professional development in this area was confirmed by two recent reports. A report on the licensing of school principals in the 50 states, sponsored by the Wallace Foundation (Adams & Copland, 2005), includes a discussion of the necessary knowledge and skills for principals, according to state licensing requirements. "Totally missing from state licensing frameworks," according to the authors, "was any attention to the meaning and use of learning assessments. . ." In another recent study, by the National Board on Educational Testing and Public Policy at the Lynch School of Education at Boston College, researchers surveyed a nationally representative sample of teachers to ascertain their attitudes about state-mandated testing programs (Pedulla et al., 2003). When asked about the adequacy of professional development in the area of standardized test interpretation, almost a third of the 4,200 responding teachers reported that professional development in this area was inadequate or very inadequate (Pedulla et al., 2003). Further documentation of the "widespread deficits in assessment skills evidenced by practicing teachers" is described by Lukin, Bandalos, Eckhout, and Mickelson (2004, pp. 26–27).

As part of the preliminary work conducted in preparation for the ITEMS project, a survey assessing respondents' understanding of educational measurement and statistics was developed and field-tested by two doctoral students under the first author's supervision (Brown & Daw, 2004). The survey, which consisted mainly of multiple-choice questions, was completed by 24 University of California, Santa Barbara (UCSB) graduate students who were pursuing an M.Ed. and a multiple-subjects teaching credential. A revised version of the survey was subsequently offered over the Internet to students enrolled in a graduate education course at California State University, Northridge (CSUN). The 10 CSUN students who responded to the Web-based version of the survey were experienced teachers or administrators in K-12 schools. Results of the two survey administrations to teachers and future teachers suggest the existence of substantial gaps in the respondents' knowledge of educational measurement and statistics. For example, only 10 of the 24 UCSB respondents were able to choose the correct definition of measurement error, and only 10 knew that a Z-score represents the distance from the mean in standard deviation units. Nearly half mistakenly thought the reliability of a test is "the correlation between student grades and student scores on the test." When told that "20 students averaged 90 on an exam, and 30 students averaged 40," only one-half of the CSUN group were able to calculate the combined average correctly, and only one in 10 chose the correct definition of measurement error.

As Popham (2006a, p. xiii) notes, "[t]oday's educational leaders need to understand the basics of assessment or they are likely to become yesterday's educational leaders . . ." How can the required level of understanding be attained? Ideally, teacher and principal certification programs will eventually be modified to increase course content in educational measurement and statistics. Substantial changes of this kind are likely to be slow, however, and may occur only if licensing requirements are first modified. To help K-12 schools respond rapidly to the training gap in these areas, the Instructional Tools in Educational Measurement and Statistics (ITEMS) for School Personnel project has created and evaluated a series of three instructional modules in educational measurement and statistics, to be available via the Web and also as CDs and DVDs.

The goal of the project was to provide immediate opportunities for teachers and administrators to increase their assessment literacy—more specifically, their understanding of the psychometric and statistical principles needed for the correct interpretation of standardized test scores. The ITEMS materials are intended to help prepare school personnel to use test results to optimize instructional decisions and to pinpoint schools, classes, or individuals that require additional instruction or resources, as well as to explain test results to students, parents, the school board, the press, and the general community. The provision of this training in a convenient and economical way is intended to assist schools with the successful implementation and interpretation of assessments. Alternative forms of professional development,

such as intensive workshops or university courses, are much more time-consuming and expensive, and are unlikely to be funded in an era of limited school budgets.

Previous Statistics Education and Assessment Literacy Research

The ITEMS work is related to previous research in statistics education and assessment literacy. The goal of statistics education research is to promote effective techniques, methods, and products for teaching statistics, as well as to understand how individuals reason about statistical concepts. By contrast, the work described here as assessment literacy research focuses on the ways teachers use assessments to make instructional decisions.

Statistics Education

A prominent statistics education research effort is that of Joan Garfield and Robert delMas of the University of Minnesota and Beth Chance of California Polytechnic State University. In the current ARTIST (Assessment Resource Tools for Improving Statistical Thinking) project (https://app.gen.umn.edu/artist/), these researchers are working to develop effective assessments of statistics knowledge, which are made available online for teachers of first courses in statistics (Garfield, delMas, & Chance, 2003; see https://app.gen.umn.edu/artist/publications.html for a publications list). In a previous project, delMas, Garfield, and Chance (1999) investigated the use of computer simulations to improve the statistical reasoning of students in college-level introductory statistics courses. These projects built on an earlier program of research, Project Chance, headed by J. Laurie Snell of Dartmouth, the goal of which was to help students think critically about media reports that use probability and statistics (Snell & Finn, 1992).

The Statistics Education Research Group (SERG) (www.umass.edu/srri/serg/) at the University of Massachusetts, Amherst works on a variety of projects to improve instruction in statistics courses. Founded in the 1970s, the group originally investigated how individuals reason about statistical concepts before receiving any formal training, and then used results to improve K-12 and college-level statistics instruction. Their current focus is on younger students. Several data analysis software tools have emerged from projects conducted at SERG, including DataScope®, Prob Sim® (Konold & Miller, 1994), and Tinkerplots™ (Konold & Miller, 2004). These products are designed for teaching concepts in probability and sampling using simulations. They have been used extensively in the classroom to teach concepts to students, and have also been used to investigate teachers' approaches to analyzing data (e.g., see Hammerman & Rubin, 2002, 2004).

While much of the statistics education literature is focused ultimately on student learning, recent studies have also examined teachers' statistical thinking and reasoning. Hammerman and Rubin (2002, 2004), Makar and Confrey (2002) and Confrey, Makar, and Kazak (2004) have examined ways that teachers reason about statistical concepts (in particular, variation in data) and how using computer graphics tools can facilitate their understanding.

Hammerman and Rubin (2002, 2004) studied the strategies that teachers used to comprehend variability in data and examined how teachers explored and analyzed data using the Tinkerplots™ software. After learning to "bin" and visualize data using the graphical tools, teachers were less likely to use the sample average as a single overall representation of the data.

Makar and Confrey (2002) examined mathematics teachers' statistical thinking about high-stakes test results at a low-performing high school in Texas. Using the computer-based statistical learning tool, *Fathom*™ (Finzer, 2001), teachers created graphs to compare distributions of test scores for male and female students to determine if the differences in the centers between the two groups of data were significant. However, the research showed that, after attending a workshop in data analysis, teachers still used intuition to determine whether differences in centers of the distributions of test scores were significant, ignoring concepts like variability and sampling distributions in their reasoning.

Confrey et al. (2004) conducted additional studies to explore the statistical understanding of high-stakes test data by teachers. The purpose was to determine if a professional development course in the area of testing and statistics could improve teacher reasoning about statistical concepts as it relates to test data. The results indicated that participants made significant gains in their understanding of statistical concepts; however, it was not determined if teachers made significant gains in their understanding of test results per se.

Assessment Literacy

Since the enactment of NCLB, an increasing number of schools and districts are collecting massive quantities of student data, including standardized test results, using software management products. (See Wayman, Stringfield, and Yakimowski (2004) for a discussion of various software products designed for storing, organizing, and analyzing student, school, and district level data.) Administrators and teachers are expected to use these data to make educational and instructional decisions. Boudett, City, and Murnane (2005) outline eight critical steps to effectively use assessments to improve instruction and raise student achievement. According to Boudett et al. (2005), the second step is to "build assessment literacy," i.e., develop a working knowledge of common concepts related to test score results, and acquire appropriate skills to interpret test score results. The text by Popham (2006a) is intended to improve the assessment literacy of educational leaders and another recent contribution by Popham (2006b) consists of 15 booklets on assessment topics that are intended to help teachers increase their assessment knowledge.

Other resources for teachers and administrators are available, although in many cases their effectiveness has not been formally assessed. Test publishing companies offer products for professional development in assessment literacy. For example, Educational Testing Service (ETS) and CTB McGraw-Hill provide written guides and workshops for professional development in which teachers can learn to use data for instructional improvement. The Pathwise® Evidence-Centered Teaching and Assessment Workshops, offered through ETS, are designed for language arts and mathematics teachers as training sessions in the principles of formative assessment and linking of assessment to instruction. The Assessment Training Institute, founded by Rick Stiggins in 1993 and acquired by ETS in 2006, provides classroom assessment training needed to support educators. Books, videos, DVDs, and seminars are available through the institute.

Project Overview

The ITEMS project differs from much of the previous work in at least three respects. First, the ultimate target audience is school personnel themselves, rather than K-12 or college students. Also, rather than focusing on either statistical reasoning (e.g., Chance, 2002 or Mills, 2002) or on classroom testing (like Stiggins & Chappuis, 2005), the project is intended to help teachers and principals become educated consumers with respect to a broad range of topics in educational measurement and statistics. Finally, while the effectiveness of many assessment products currently available has not been evaluated, a major research component of the ITEMS project has been dedicated to investigating the effectiveness of the modules.

The project work was organized so that one module was produced and evaluated in each of the three years of the project. The topics of the three modules are as follows:

Module 1: Test Scores and Score Distributions
Module 2: Imprecision in Individual and Average Test Scores
Module 3: Interpretation of Test Score Differences and Trends

A brief outline of the content of the three modules appears in Table 1.

For each module, the project work consisted of the following four phases, which took place over a period of roughly one year:

Development phase. The module is conceptualized and produced.

Research phase. Data are collected and analyzed to allow formal evaluation of the effectiveness of a Web-based version of the module.

Program evaluation phase. The project evaluator collects data from participants to determine their views on the usefulness and effectiveness of the module.

Dissemination phase. The module is posted on our project Website as a freely available resource for educators, along with supplementary materials. CD and DVD versions of the module are mailed to those who request them.

Principles of Module Development

The instructional modules rely on a case-based approach to learning (Lundeberg, Levin, & Harrington, 1999) in which realistic test score reports are used as a basis for explaining concepts and terminology. In each module, features of these reports are highlighted and explained. In computer-based learning environments, it has been found that individuals who are presented with material via an animated agent demonstrate better learning outcomes than those who are presented with the material via on-screen text and static graphs (Moreno, Mayer, Spires, & Lester, 2001). Therefore, the modules make liberal use of graphics, including computer animation, to present concepts.

For example, Figure 1 shows a screen shot from our second module that illustrates the concept of measurement error. The viewer is asked to imagine that Edgar, the child pictured in the display, takes a test several times, magically forgetting the content of the test between administrations. On the first occasion, he misreads a question; on the second, he guesses correctly; and on the third, he is accidentally given extra time. For these reasons, he gets slightly different scores on each imaginary test administration.

In designing the modules, we have incorporated other findings from the cognitive psychology literature on learning through technology, as reflected in the following principles:

Multimedia principle
Concepts are presented using both words and pictures (static or dynamic graphics). Research has indicated that "...human understanding occurs when learners are able to mentally integrate visual and verbal representations" (Mayer, 2001, p. 5).

Contiguity principle
Auditory and visual materials on the same topic are, whenever possible, presented simultaneously, rather than successively, and words and corresponding pictures appear on the screen together rather than separately. Materials that incorporate these principles of temporal and spatial contiguity have been shown to enhance learning (Mayer, 2001, pp. 81–112).

Modality principle
Verbal information that accompanies graphical presentations, is, in general, presented in spoken form rather than as on-screen text. Research has indicated that spoken verbal material is preferable in this context because, unlike displayed text, it "does not compete with

Table 1. Content of Instructional Modules

1. "What's the Score?" (2005): Test Scores and Score Distributions
- Mean, median, mode
- Symmetric vs. skewed distributions
- Range, standard deviation
- Percentage above a cut-point (NCLB)

- Raw scores
- Percentile ranks
- Grade-equivalents
- Norm-referenced and criterion-referenced score interpretation

2. "What Test Scores Do and Don't Tell Us" (2006): Imprecision in Individual and Average Test Scores
- Measurement error and test reliability
- Standard error of measurement
- Confidence bands for test scores
- Effect of sample size on precision of test score means
- Precision of individual scores versus precision of means
- Test bias

3. "What's the Difference?" (2007): Interpretation of Test Score Differences and Trends
- Importance of disaggregating data for key subgroups
- Effect of population changes on interpretation of trends
- Effect of number of students
- Effect of changes in tests and test forms; test equating

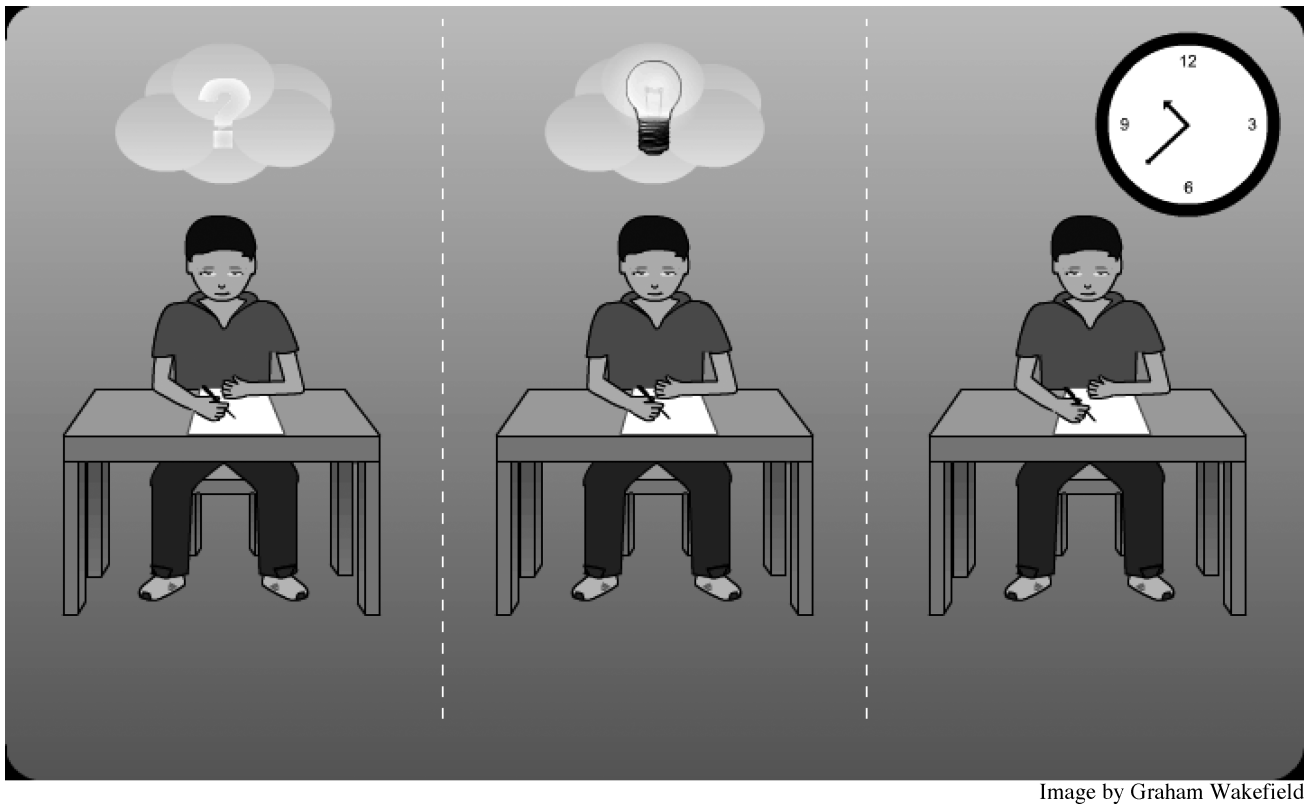

Image by Graham Wakefield

FIGURE 1. Illustration of measurement error from Module 2.

pictures for cognitive resources in the visual channel" (Mayer, 2001, p. 140).

Coherence principle

Efforts have been made to remove extraneous words, pictures, and sounds from the instructional presentations. Irrelevant material, however interesting, has been shown to interfere with the learning process (Mayer, 2001, pp. 113–133).

Prior knowledge principle

The modules are designed to "use words and pictures that help users invoke and connect their prior knowledge" to the content of the materials (Narayanan & Hegarty, 2002, p. 310). For example, while participants may be unfamiliar with the term "sampling error," most will be familiar with statements like, "The results of this survey are accurate within plus or minus five percentage points." Analogies and metaphors have also been shown to enhance mathematical learning (English, 1997).

Conversational style

A conversational rather than a formal style is used in the modules; this has been shown to enhance learning, perhaps because "learners may be more willing to accept that they are in a human-to-human conversation including all the conventions of trying hard to understand what the other person is saying" (Mayer, 2003, p. 135). In keeping with this principle, formulas are not used in the instructional modules. (The relevant formulas are given in the supplementary handbooks for those who are interested in learning them.)

Project Staffing

The project staff who created the three instructional modules described in this article consisted of six individuals (some of whom worked on only one or two modules): The principal investigator (first author), who is an education professor with over 25 years of research experience in educational measurement and statistics; a senior researcher (second author), who is a statistics professor with experience in education research; three technical specialists (third, fifth, and sixth authors) who were graduate students in media arts and technology with experience in computer programming, graphics, and animation, and a professional animator (fourth author). The project staff also included a project administrator, who was responsible for much of the day-to-day management of the project and a project evaluator, who conducted a semi-independent review of the project's effectiveness.

The project staff met three times a year with an advisory committee consisting of public school teachers and administrators and university experts in human-computer interaction, multimedia production, multimedia learning, cognitive psychology, teacher education, educational technology, theoretical statistics, and math and statistics education.

Principles of Program Evaluation

Generally stated, the questions to be addressed by the program evaluation procedures for the ITEMS project are consistent with those considered by Owen (2007, p. 48) under the rubric of impact evaluation: "Have the stated goals of the program been achieved? Have the needs of those served by the program been achieved? . . . Is the program more effective for some participants than for others?" In our case, the stated goal of the project was to present psychometric and statistical information about the interpretation of standardized test scores that

was comprehensible and applicable to everyday work situations encountered by the participants. Furthermore, we hoped that the information would be retained. In addition, we wanted the presentation format to be perceived as appealing and convenient by the users.

Four sources of information were used to address the evaluation questions. First, as described below, a quiz was constructed to correspond to each module. The purpose of the quiz was to determine whether participants comprehended the information presented in the module and could apply this knowledge to problems intended to resemble those encountered in their everyday work. An experimental paradigm, involving random assignment was implemented in which half the participants took the quiz before seeing the module, and half took the quiz after viewing the module. Participants willing to be followed up took the quiz a second time, one month later, to measure retention. Second, a background survey was used to collect information about participants so that we could determine if the modules were more useful and effective for some participants than for others. Third, an independent evaluator conducted phone interviews (Module 1) and administered an evaluation survey (Modules 1, 2, and 3) to participants to assess the utility and effectiveness of the materials and to solicit suggestions for improvement. (The evaluator role was filled by one individual for the Module 1 evaluation and a second individual for the evaluation of Modules 2 and 3.) Fourth, participants provided comments via comment boxes included as part of the module and quiz administration, and, in some cases, via email messages to the project.

Development and Evaluation of Module 1, "What's the Score?"

Module 1 Development

The development of our first module, "What's the Score?" began in 2004. By displaying conversations between the cartoon characters Stan, a somewhat clueless teacher, and Norma, a more informed one, this 25-minute module explores topics such as test score distributions and their properties (mean, median, mode, range, standard deviation), types of test scores (raw scores, percentiles, scaled scores, and grade-equivalents), and norm-referenced and criterion-referenced score interpretation. Table 1 provides some additional information on the content of the module.

The module development process began with a wide-ranging review of literature, including research articles and software products related to statistics education and assessment literacy training, textbooks in measurement and statistics, the *Standards for Teacher Competence in Educational Assessment of Students* (American Federation of Teachers, NCME, & National Education Association, 1990), the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & NCME, 1999), research findings on the effectiveness of multimedia instruction (see above), and references on storyboarding and film-making (e.g., Begleiter, 2001). The module development process also drew on the results of a preliminary survey of the measurement and statistics knowledge of teachers, school administrators, and teacher education students.

Following these preliminary steps, a detailed outline of the module was developed. Text was then created and scenes were conceptualized and incorporated in a storyboard. To maximize accessibility, both low- and high-bandwidth versions of the Web-based module were produced using Macromedia Flash® (now Adobe Flash) software. The module was revised over a period of six months to reflect the feedback received. For example, after viewing an initial version, our advisory committee recommended shortening the module and dividing it into short segments, each preceded by a title slide. This recommendation was implemented before the research phase began. Also, on the recommendations of participants, "pause" and "rewind" capabilities were added to allow viewers to navigate among the seven segments of the module.

Module 1 Research Phase

The primary means of evaluating the effectiveness of the module involved the administration of a Web-based assessment literacy quiz that we developed to address the topics covered in the module. A 20-item multiple-choice quiz was pilot-tested on a group of teacher education students at UCSB and then revised substantially to improve its correspondence to the material included in the module. The purpose of the quiz was to allow us to determine whether viewing that module improved understanding of the concepts that were presented.

Sixty-eight teacher education program (TEP) students from UCSB, as well as 45 teachers and administrators from local school districts, were recruited to take part in a formal study of the effectiveness of the Web-based module. Participants received $15 gift cards from Borders. Demographic information on the participants is given in Table 2. Overall, participants had an average age of 34 and had taught for an average of seven years. Seventy percent were women. The TEP students, of course, tended to be younger and have fewer years of teaching experience than the school personnel.

When they logged on, participants completed a background survey and then were randomly assigned to one of two conditions: In one condition, the module was viewed before the quiz was administered; in the other, the quiz was administered first. By comparing participants from the two conditions, we were able to test the hypothesis

Table 2. Module 1: Demographic Information for Participants

	Sample Size	Average Age	Percent Female	Percent in Administrative Positions	Average Years of Teaching Experience
All Participants	113	33.7	69.9	5.3	7.4
Teacher Education (TEP) Students	68	25.4	75.0	0.0	2.3
School District Personnel	45	46.6	62.2	13.3	15.1

Table 3. Module 1: Results on the 20-Item Quiz

	Module-First Group			Quiz-First Group			Effect Size	*t*-test *p*-value (1-sided)
	Mean	*SD*	*n*	Mean	*SD*	*n*		
All Participants	13.2	3.7	52	12.0	3.4	61	.34	.042
Teacher Education (TEP) Students	13.1	4.0	33	11.7	3.5	35	.37	.059
School District Personnel	13.4	3.2	19	12.5	3.2	26	.28	.198

that those who viewed the module first were better able to answer the quiz questions.

Data analysis

Psychometric analysis of the assessment literacy quiz showed that, for the total group, the quiz had an average percent correct of 63, an average item discrimination (corrected item-total correlation) of .28, a reliability of .71, and a standard error of measurement of 1.9 points.

As indicated in Table 3, results for the 113 participants showed that viewing the module had a small but statistically significant positive effect on quiz scores: Those in the "module-first" group answered an average of 13.2 of 20 questions correctly, compared with an average score of 12.0 questions for the "quiz-first" group (an effect size of .34 standard deviation units).

Further analysis showed that the quiz item on which the two groups differed most, addressed the topic of skewness of a test score distribution. On this question, 63% of the module-first group answered the question correctly, while only 34% of the quiz-first group answered the question correctly. On the remaining quiz items, differences in the percentages correct for the two groups were typically less than 10 points.

When the overall results in Table 3 are disaggregated, it becomes apparent that the effect was larger among the TEP students than among the school personnel (see Figure 2). There are several possible explanations for this disparity. As noted, the TEP students differed in several ways from the school personnel (Table 2). In addition, Module 1 and the associated quiz were presented to these students in a group administration, with project staff available to trouble-shoot. By contrast, school personnel participated on an individual basis, at a time and place of their own choice.

In addition to comparing the quiz results by TEP status, we were also interested in comparing quiz results by math instructor status. Participants who were either full-time math teachers or TEP student teachers were classified as math instructors if they indicated that math was the sole subject taught, while all other participants were classified as nonmath instructors. Because we expected math instructors to be more familiar with measurement and statistics concepts than other participants, we anticipated that they would perform better on the quiz than other participants. In fact, math instructors ($n = 19$), scored only slightly higher than participants who were not math instructors ($n = 94$). The average score for math instructors was 13.2 points ($SD = 2.9$), compared to an average score of 12.4 for other participants ($SD = 3.7$).

Follow-up quiz analysis

A follow-up analysis was conducted to determine the extent to which participants retained an understanding of the module material, as measured by their follow-up performance on the quiz. Individuals who agreed to participate in the follow-up phase of the Module 1 research were contacted about a month after they had first viewed the module and taken the quiz. They were given an online quiz identical to the one that they took during their first module-viewing session.

Unfortunately, of the original 113 participants, only 11 participated in the follow-up phase and completed the quiz again. Four participants were from the module-first group, and seven were from the quiz-first group. The average score on the quiz for the four members of the module-first group was 15.5 the first time they took the quiz, and it remained 15.5 when they retook the quiz (note that there was variability within the scores at the two different times). The average quiz score for the seven quiz-first follow-up participants was 14.3 the first time they took the quiz, and increased to 15.9 when they took it during the follow-up phase. Although the follow-up samples are small

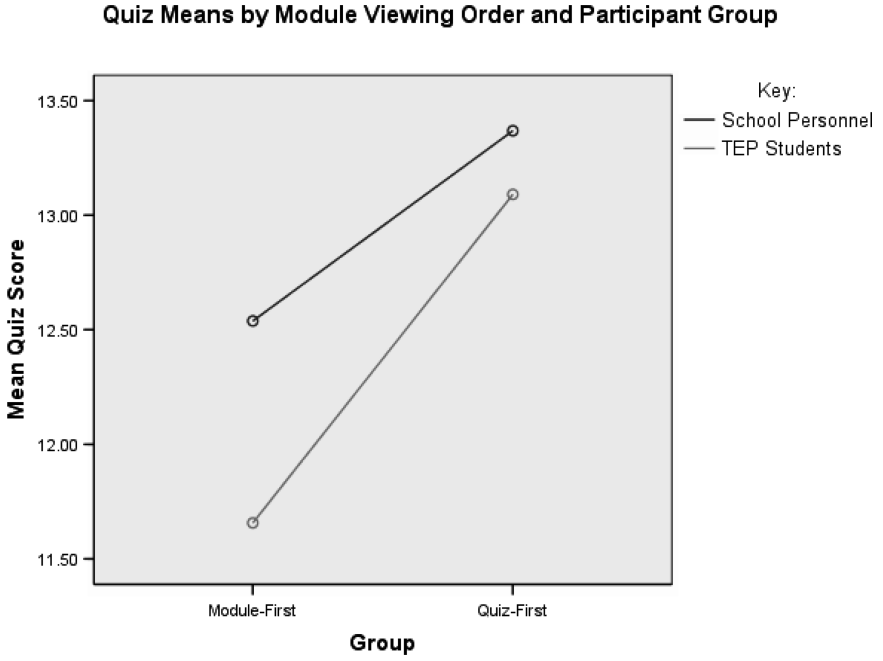

FIGURE 2. Module 1: Average scores on 20-item quiz for school personnel and teacher education program students in quiz-first and module-first groups.

and cannot be assumed to be representative of the original participants, it is worth noting that on average, no loss of content knowledge was observed for either group.

Development and Evaluation of Module 2, "What Test Scores Do and Don't Tell Us"

Module 2 Development

Our second module, "What Test Scores Do and Don't Tell Us," shows Stan, the teacher introduced in Module 1, meeting with parents Maria and Tim to discuss the test results of their twins, Edgar and Mandy. The module focuses on the effect of measurement error on individual student test scores, the effect of sample size on the precision of average scores for groups of students, and the definition and effect of test bias. Table 1 provides some additional information on the content of the module.

As was the case in developing Module 1, a detailed outline was first developed. Text was then created and scenes were conceptualized and incorporated in a storyboard. Following that, a Web-based version of the module was produced using Flash® software and was revised over a period of several months to reflect the feedback received from our project advisory committee. Both low- and high-bandwidth versions were made available.

Module 2 includes two features not incorporated in Module 1. First, an optional closed captioning feature is available, which should facilitate the use of the module by participants who are hard of hearing. Also, the module includes four "embedded questions" (one for each main section) that allow participants to check their understanding of the material and view portions of the module a second time if they wish. The embedded questions are also intended to encourage participants to remain engaged in the process of watching the module.

Recruitment and Data Collection Procedures for Module 2

The 2006 research and program evaluation phases differed from their 2005 counterparts in three related respects. First, whereas we recruited only Central California teachers and educators in 2005, we recruited educators from around the nation in 2006. Our primary means of doing so was to place an advertisement in the educational magazine *Phi Delta Kappan* in January and February. We also adapted our Website to allow individuals to sign up for the project on the Website itself. In addition, we contacted professional organizations, focusing on those for minority professionals, to encourage them to sign up for project participation.

A second change that we implemented in 2006 was necessitated by the fact that we were working with participants who were in some cases thousands of miles away, making visits to schools impractical. We therefore developed an infrastructure that allowed all transactions that take place during the research and program evaluation phases to be conducted electronically. Provision of logon information to participants, collection of data, and even distribution of gift cards were electronic. (Participants received electronic certificates that are usable at Borders.com.)

A third change was that, rather than regarding a school or district as a target of recruitment, we directly enlisted the participation of individual teachers and administrators. This did not preclude the possibility that particular schools and districts would participate; rather, it allowed us to collect data from educators whether or not their schools or districts were participating on an institutional level.

Module 2 Research Phase

To test the effectiveness of Module 2, we used a new assessment literacy quiz, tailored to the topics of Module 2. Consistent with the recommendations of our advisory panel, we created an instrument that is more applications-oriented than the Module 1 quiz. Most questions include tables or graphics that resemble those that appear in standardized test results. Respondents are asked questions about the interpretation of these displays. Because the quiz was more directly linked to the material in the module than was the case for Module 1, we believe the Module 2 quiz provided a better test of the effectiveness of the module. This 16-item multiple choice quiz was pilot-tested with UCSB TEP students.

One hundred four individuals participated in the research phase for Module 2, "What Test Scores Do and Don't Tell Us." Of those, 81 were TEP students from at least five universities, and 23 were school district personnel from 19 school districts across the country. Participants received $15 electronic gift certificates from Borders.com. Demographic information on the participants for this phase is given in Table 4. Sixty-eight percent of the participants were women, with average age of 31 years and average teaching experience of 11 years. The TEP students were typically younger and had fewer years of teaching experience than the school personnel.

To develop a better picture of the school districts where participants worked, either full-time or as a student teacher, we collected additional information about the community surrounding the school district and the minority composition of the school district. In the background survey, participants were asked if they would describe the community surrounding the school district as "central city," "urban fringe/large town," or "rural/small town." Of the 85 participants who responded to this question, 69% described the area surrounding their district as urban fringe, while 18% indicated that their community was rural, and 13% described it as central city.

Participants were also asked to estimate a range for the percentage of students in their school district who were African-American, Hispanic/Latino, Native American, or members of other ethnic minorities. The choices were as follows: less than 20%, 21%–40%, 41%–60%, 61%–80%, and 81%–100%. Diverse districts were well represented in the study. Of the 84 participants who responded to the question, 45% indicated

Table 4. Module 2: Demographic Information for Participants

	Sample Size	Average Age	Percent Female	Average Years of Teaching Experience
All Participants	104	31.4	68	11.3
Teacher Education (TEP) Students	81	25.7	73	2.6
School District Personnel	23	50.9	52	23.0

Table 5. Module 2: Results on the 16-Item Quiz

	Module-First Group			Quiz-First Group			Effect Size	*t*-test *p*-value (1-sided)
	Mean	*SD*	*n*	Mean	*SD*	*n*		
All Participants	12.6	3.0	51	10.2	3.5	53	.74	.000
Teacher Education (TEP) Students	12.6	3.2	40	9.5	3.7	41	.90	.000
School District Personnel	12.7	1.9	11	12.5	1.4	12	.12	.375

that they worked in a district composed of 41% to 60% minority students, 19% worked in a district that was 61% to 80% minority students, 13% worked in a district that was less than 20% minority students, 12% worked in a district that was 81% to 100% minority students, and 11% worked in a district that consisted of 21% to 40% minority students.

Psychometric analysis of the 16-item multiple-choice assessment literacy quiz showed that the quiz had an average percent correct of 71, an average item discrimination of .40, a reliability of .79, and a standard error of measurement of 1.58 points.

As indicated in Table 5, results for the 104 participants showed a statistically significant effect: Those in the module-first group answered an average of 12.6 out of 16 questions correctly, compared with an average score of 10.2 questions for the quiz-first group (an effect size of .74).

Further analysis of the quiz results showed that the items that had the largest group differences in the percentages correct—between 20% and 27%—were primarily about measurement error or about the relation between the stability of sample means and sample size. For example, a question about the standard error of measurement was answered correctly by 88% of the module-first group, compared to 62% of the quiz-first group. For the remainder of the questions, the typical difference between the groups in the percentages correct was between 5% and 10%.

Table 5 shows that, as in Module 1, the effect of viewing the module was much larger for the TEP students (an effect size of .90) than for the school personnel (.12), a finding that is discussed further below. In Figure 3 we can observe that school personnel performed about the same (on average) on the quiz regardless of whether they viewed the module before or after taking the quiz; however, there was a difference in the average scores between TEP students who viewed the module before and after taking the quiz. This effect is more apparent than that obtained in the Module 1 evaluation phase.

To parallel the analysis in the Module 1 evaluation phase, we compared quiz results between math instructors and nonmath instructors. Math instructors consisted of both full-time math teachers and TEP student teachers in math; all other participants were classified as nonmath instructors. As in the Module 1 evaluation phase, math instructors ($n = 5$) scored slightly higher on average than participants who were nonmath instructors ($n = 99$). The average score for math instructors was 14.4 ($SD = 1.5$), compared to 11.2 for nonmath instructors ($SD = 3.5$). (Note that only five participants were designated as math instructors in the Module 2 analysis, as compared to 19 in the Module 1 analysis. This may be because information on teaching was collected slightly differently on the two occasions. In the Module 1 background survey, participants were asked to indicate the subject they taught most frequently, while in the Module 2 survey, they were given the option to indicate multiple subjects that they taught. Participants who indicated only "math" were then designated as math instructors.)

Follow-up quiz analysis

As in the first module research phase, a follow-up analysis was conducted to determine the extent to which participants retained an understanding of the material from the second module. Participants in the Module 2 research phase were contacted about a month after they had first viewed the module and taken the quiz. They were given the identical quiz that they had previously taken during their first module-viewing session.

Due to improvements in the online administration system, contact with participants was more easily achieved,

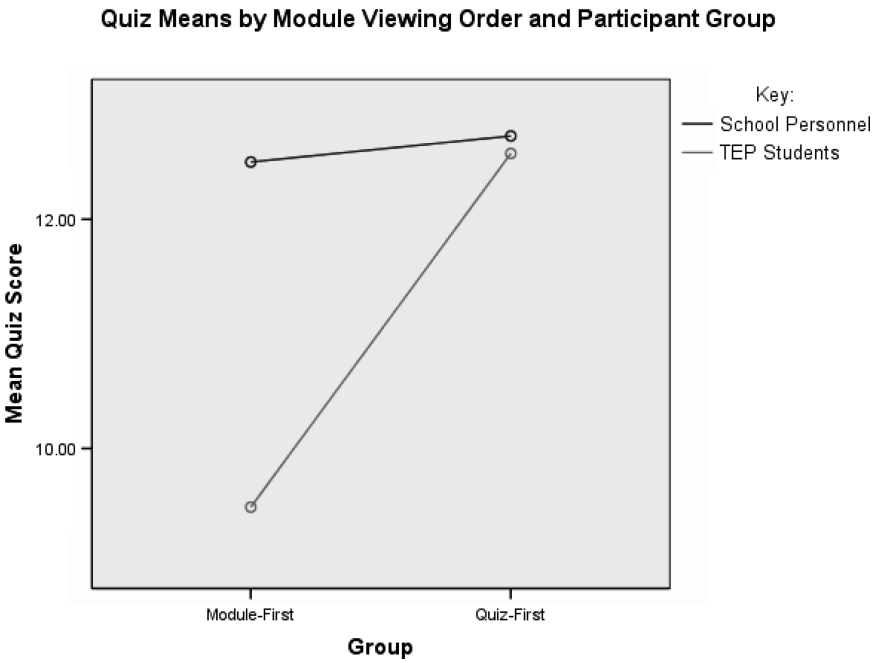

FIGURE 3. Module 2: Average scores on 16-item quiz for school personnel and teacher education program students in quiz-first and module-first groups.

Table 6. Module 3: Demographic Information for Participants

	Sample Size	Average Age	Percent Female	Average Years of Teaching Experience
All Participants	33	37.3	82	6.6
Teacher Education (TEP) Students	14	25	93	1.0
School District Personnel	19	46.4	74	13.8

and a much higher follow-up rate was observed. Of the original 104 participants who took the quiz, 38 participated in the follow-up phase and completed the quiz again. Fifteen participants were from the module-first group, and 23 were from the quiz-first group. The average score on the quiz for the 15 members of the module-first group was 13.9 the first time they took the quiz, and it dropped slightly to 13.1 when they retook the quiz. However, this difference was not statistically significant ($p = .20$) at the .05 level, as determined by a matched-pairs t-test. Hence, there does not appear to be any significant loss of content knowledge for this group. The average quiz score for the 23 quiz-first follow-up participants was 10.6 the first time they took the quiz, and increased to 11.4 when they took it during the follow-up phase. This difference was marginally statistically significant ($p = .05$), which is consistent with an increase in content knowledge due to the effect of viewing the module. As in the Module 1 evaluation, interpretation is complicated by the fact that the follow-up samples are small and cannot be assumed to be representative of the original participants.

Development and Evaluation of Module 3, "What's the Difference?"

Module 3 Development

Our third module, "What's the Difference?" shows a press conference in which a superintendent is explaining recently released test results. Norma and Stan, the teachers introduced in the previous modules, help to demystify the test results for the reporters. The main topics addressed are the importance of disaggregating test data for key student groups and the implications for score trend interpretation of shifts in the student population, the number of students assessed, and changes in tests and test forms. (See Table 1 for details.)

As was the case in developing Modules 1 and 2, an outline and script were first developed. An animated Web-based module was then produced, revised, and incorporated into our data collection system. The closed captioning and embedded questions features introduced in Module 2 were retained.

Production values for Module 3 were substantially improved over the previous two modules. First, a professional animator with extensive cartooning experience joined the project. Second, students from the Dramatic Art department at UCSB were hired to voice the animated characters, rather than employing project staff for this purpose. Finally, the audio recording was conducted in a professional sound studio. These changes resulted in a module that was much more polished than its predecessors.

Recruitment and data collection procedures for Module 3 were similar to those in Module 2, except that we selected a different magazine—*Learning and Leading with Technology*—in which to place our advertisement, and we timed the ad to be concurrent with the data collection instead of preceding it. As an added incentive to participation, a feature was added that allowed participants the option of downloading a personalized certificate indicating that they had completed an ITEMS training module.

Module 3 Research Phase

The effectiveness of Module 3, "What's the Difference?" was tested using an assessment literacy quiz tailored to the topics of the module. We began with a 16-item instrument, which was reduced to 14 items after being pilot-tested with UCSB TEP students.

Primarily because of reduced participation by the UCSB TEP students during the research phase, we were initially able to recruit only 23 participants. Four were UCSB TEP students and 19 were school district personnel from seven states. We were subsequently able to collect data from 10 teacher education students at California State University, Fresno, bringing the number of TEP students to 14. As before, participants received $15 electronic gift certificates from Borders.com.

Demographic information on the participants for this phase is given in Table 6. About 82% of the participants were women, the average age was 37, and the average number of years of teaching experience was 6.6. The teacher education students were generally younger and had fewer years of teaching experience than the school personnel.

As in the Module 2 research phase, we collected information about the community surrounding the school district and the minority composition of the school district. Participants were asked if they would describe the community surrounding the school district as "central city," "urban fringe/large town," or "rural/small town." Of the 24 participants who responded to this question, 12 described the area surrounding their district as urban fringe, 3 indicated that their community was rural, and 9 described it as central city.

Participants were also asked to estimate a range for the percentage of students in their school district who were African-American, Hispanic/Latino, Native American, or members of other ethnic minorities. The choices were identical to those given in the Module 2 background survey: Less than 20%, 21%–40%, 41%–60%, 61%–80%, and 81%–100%. Diverse districts were well-represented: Of the 24 participants who responded to the question, 8 worked in a district that consisted of 21% to 40% minority students, 7 indicated that they worked in a district composed of 41% to 60% minority students, 4 worked in a district that was less than 20% minority, 3 worked in a district that was 81% to 100% minority, and 2 worked in a district that was 61% to 80% minority.

Psychometric analysis of the 14-item multiple-choice assessment literacy quiz revealed that the quiz had an average percent correct of 63, an average item discrimination of .53, a reliability of .87, and a standard error of measurement of 1.47 points.

Table 7. Module 3: Results on the 14-Item Quiz

	Module-First Group			Quiz-First Group			Effect Size	*t*-test *p*-value (1-sided)
	Mean	*SD*	*n*	Mean	*SD*	*n*		
All Participants	9.1	4.2	18	8.5	4.1	15	.14	.33
Teacher Education (TEP) Students	6.5	4.1	8	5.5	2.1	6	.32	—
School District Personnel	11.2	3.0	10	10.4	4.0	9	.23	—

Note: Because of small sample sizes, a single *t*-test was performed for all participants combined; separate tests were not conducted for TEP students and school district personnel.

As indicated in Table 7, the effect of the module was not statistically significant. (Because of small sample sizes, a single *t*-test was performed for all participants combined; separate tests were not conducted for TEP students and school district personnel). Those in the module-first group did have a slightly higher average quiz score (9.1) than those in the quiz-first group (8.6), however (see Figure 4). In both the module-first and quiz-first groups, school personnel performed substantially better than teacher education students. There were not enough data to make a meaningful comparison between math and nonmath instructors as was done for the first two research phases. Only two participants were math instructors. One math instructor with three years of teaching experience received a perfect score on the quiz, while the other math instructor, who is currently a TEP student, answered 5 out of 14 items correctly.

Further analysis of the quiz results did not show much difference between the scores of the module-first and quiz-first groups on most items. However, two questions did reveal substantial group differences: An item on the effect of sample size on the interpretation of improvements in average test scores was answered correctly by 83% of the module-first group, compared to 60% of the quiz-first group. A question about test equating was answered correctly by 78% of the module-first group, compared to only 27% of the quiz-first group.

Follow-up quiz analysis

A follow-up analysis was conducted to determine the extent to which participants retained an understanding of the material from the second module. Participants in the Module 3 research phase were contacted about a month after they had first viewed the module and taken the quiz. They were given the identical quiz that they had previously taken during their first module-viewing session. Of the original 33 participants who took the quiz, 11 (33%) participated in the follow-up phase and completed the quiz again. Four participants were from the module-first group and seven were from the quiz-first group. The average score on the quiz for the four members of the module-first group was 12.3 the first time they took the quiz, and it remained the same when they retook the quiz. The average quiz score for the seven quiz-first follow-up participants was 10.7 the first time they took the quiz, and increased to 12.6 when they took it during the follow-up phase. Because of the very small number of participants in the follow-up phase, tests of significance were not conducted. As in the results from the previous follow-up studies, the slight increase in average score is consistent with an increase in content knowledge due to the effect of viewing the module.

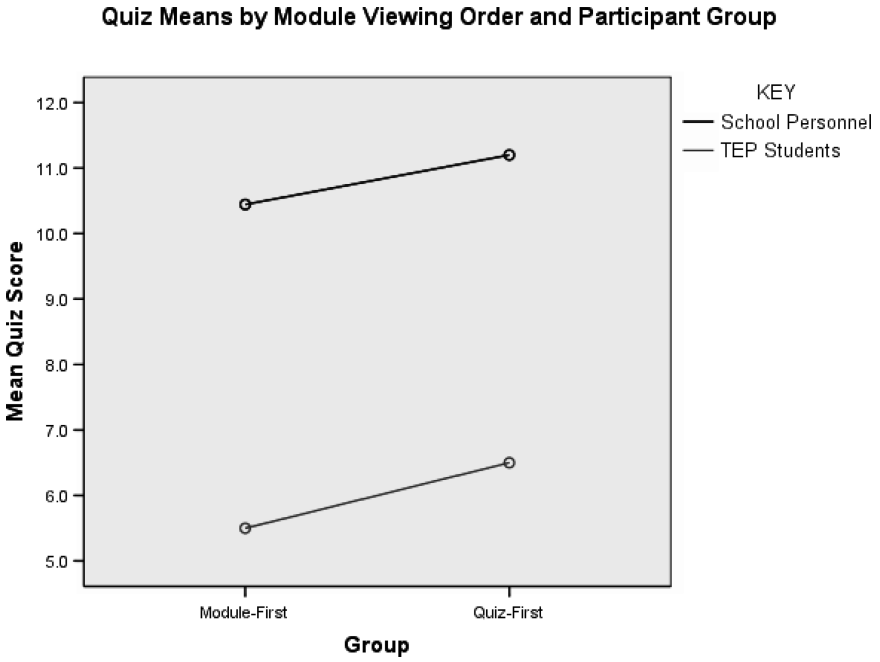

FIGURE 4. Module 3: Average scores on 14-item quiz for school personnel and teacher education program students in quiz-first and module-first groups.

Program Evaluation and Dissemination Phases for Modules 1, 2, and 3

During the program evaluation phases, project participants were asked to provide data to the project evaluator about the quality and usefulness of the modules. Requests to participate in the evaluation came directly from the evaluator, rather than the project director, and no incentives were provided for participation in this phase. Our intention was to keep the program evaluation phase as separate as possible from the main activities of the project, in hopes that participants would feel free to provide honest comments on the

project materials. The response rate for the program evaluation phase was low in all three years. For Module 1, seven teachers and six administrators (11.5% of the original participants) agreed to participate in phone interviews or complete paper surveys administered in person or by mail by the evaluator. In the survey, participants were asked to rate several aspects of the project on a 4-point Likert scale (poor, fair, good, excellent). Questions about the overall program, comprehensiveness of the material and presentation of material yielded modal responses of "good"; questions on "relevance to my role" and availability and access yielded modal responses of "excellent." In addition, most participants said they had used or planned to use the materials in their work. Interview comments were "very positive, in general" according to the evaluation report, but included both complimentary and critical feedback about the content and technical features of the module. The more negative comments tended to focus on the unavailability of navigation features, a problem that was later corrected.

The Module 2 program evaluation phase included a 22-item evaluation survey that was administered online. Approximately 5 weeks after their original participation, individuals received an email invitation to participate in the evaluation. Eleven individuals (10.6%) completed the surveys, which contained Likert items about the presentation, content, and impact of the module and open-ended questions about the quality of the module. Responses regarding presentation (e.g., quality of navigational tools) were uniformly positive, and responses on content (e.g., quality of examples) were all positive as well, except for one response to one of the five content questions. In the impact section, most respondents reported that they learned from the module, increased their confidence in the subject matter, and expected to change the way they talked to others about test results, though there was some disagreement among respondents on these issues. The open-ended questions on quality yielded positive responses about the clarity of presentation and about the examples and terms included in the module, as well as some negative responses about animation quality and about specific aspects of the content. Informal comments that participants entered in comment windows during the original or follow-up data collection were also analyzed. Roughly one-third of the overall comments made were positive, one third were neutral, and one-third were negative. Among the recommendations made by the participants were to make definitions of terms available in the module through pop-up windows or some other means, and to increase the number of examples used to illustrate concepts.

The Module 3 evaluation procedures and survey paralleled those of Module 2. Unfortunately, only two of the original 33 participants (6.1%) completed the survey. These two respondents provided uniformly positive responses about the presentation, content, and impact of the module. Nineteen participants made a total of 27 comments in the windows provided during the module and quiz phases. Eleven of these comments were positive, 10 were negative, and 6 were neutral. The comments lacked a common theme. For example, while two viewers found the video aspect of the module distracting, another praised the graphics. And while one viewer thought the video was slow and repetitive in its treatment of the material, another called it "somewhat advanced."

In addition to the information collected via the evaluation survey and comments boxes, we also asked Module 3 participants to respond, following the module, to a multiple-choice item soliciting their views about the embedded questions posed at the end of each scene. Of the 32 participants who responded, 30 found the questions "somewhat helpful" or "very helpful," while 2 other participants indicated they were "neither annoying nor helpful." None considered them "distracting or annoying."

Following their respective research and program evaluation phases, the three modules were revised according to the recommendations of participants and advisory committee members. They were then made freely available on the Web, along with their associated quizzes and online "handbooks" that provide formulas, supplementary explanations, and references. Educators who preferred not to use the Web-based version of the module could request CDs or DVDs, which were mailed to them at no cost.

Discussion

A key finding of this study is that our instructional modules are effective training tools for teacher education program students. TEP students who took the quiz after seeing the module performed better than those who took the quiz before seeing the module. The effect sizes for these students were .37 for Module 1 (Table 3), .90 for Module 2 (Table 5), and .32 for Module 3 (Table 7). Results were statistically significant only for the first two modules. Follow-up analyses suggested that this information was retained a month later; however, because the small number of individuals who participated in the follow-ups cannot be assumed to be representative of the total group of initial participants, this finding must be regarded as tentative.

In the case of school personnel, there was no statistically significant difference between those who saw the module first and those who took the quiz first (Tables 3, 5, and 7). For all three modules, school personnel, on average, outperformed TEP students. For Module 2, which produced the largest effect for TEP students, the average quiz score for the school personnel (12.7 and 12.5 for the module-first and quiz-first groups, respectively) was nearly identical to that of the TEP students who saw the module before taking the quiz (12.6). This suggests that, whether or not they had seen the module, the school personnel, who had an average of 11 years of experience, were as familiar with the included measurement and statistics material as the TEP students who had already viewed the module.

A factor that somewhat complicates the interpretation of the larger effects for the teacher education students for all three modules is that, one to four months before the research phase, a portion of them had participated in a pilot test of the assessment literacy quiz. Those who took part in the pilot test responded to the quiz and provided comments about the clarity of the wording and the difficulty of the material. They did not receive the answers to the quiz, nor did they view the module. Those who took part in the pilot were just as likely to end up in the module-first group as in the quiz-first group during the research phase, so that any effects of pilot participation should have affected both experimental groups equally. Furthermore, in the case of the Module 1 quiz, the

instrument was almost entirely overhauled before the research phase, so that the final version bore little resemblance to the pilot version. Nevertheless, it is possible that pilot participation had some effect.

More generally, caution is warranted when interpreting the data analysis results. While the TEP students from UCSB were required by their own faculty to participate in the main data collection for Modules 1 and 2, the remainder of the individuals who participated in the research phase chose to do so, resulting in nonrandom samples of the corresponding populations. In addition, samples ranged from small to moderate in size and cannot be assumed to be representative of school personnel or teacher education program students nationwide. Nevertheless, the results are encouraging and do indicate that overall, the modules are having a positive impact on content knowledge in educational measurement and statistics.

Supplemental analysis of quiz performance showed that math instructors tended to perform better than other participants (Modules 1 and 2) and that the topics participants were most likely to learn about from the modules were skewness of a test score distribution (Module 1), measurement error (Module 2), the stability of sample means (Module 2), the effect of sample size on the interpretation of changes in average scores (Module 3), and test equating (Module 3).

Most comments relayed to the project via email (not included in the formal program evaluation) have been positive. Some examples are as follows:

> "Very helpful and right to the point. If I were a building principal or a department chair today all of the staff would go through this until everyone really understood it."

> "I am inclined to recommend [Module 1] as required viewing for all new hires in our K-12 district, and it certainly will be recommended . . . for inclusion in professional development on assessment literacy."

> "I will be sharing [Module 1] with my Assistant Superintendent with the hope of promoting it as a part of our new teacher induction process."

In addition, the teacher education programs at UCSB and at California State University, Fresno have now incorporated the ITEMS materials into their curriculums.

By far the greatest challenge in the ITEMS project has been the recruitment of participants. Overworked teachers who are overwhelmed by the demands of NCLB are unlikely to undertake an optional activity like this one. Ironically, then, one reason that educators do not have time to learn about the technical aspects of testing is that they are occupied with the preparation and administration of tests.

As part of our effort to increase awareness of the project, we have issued press releases and made contacts with the Corporation for Educational Network Initiatives in California (CENIC), the University of California Office of the President, the California Teachers Association Institute for Teaching, the California Department of Education Beginning Teacher Support and Assessment program, and the California County Superintendents Educational Services Association. In addition, we have made conference presentations, posted project information on education-oriented Websites and blogs, and used listservs to contact professional organizations, focusing on those for minority professionals.

Following the completion of the supplementary data collection, we will focus for the remainder of the project on the dissemination of our materials, particularly to teacher education programs and new teachers. It is our hope that even those educators who were hesitant to participate in the research aspects of the project will find the ITEMS materials useful as a resource.

Acknowledgments

We appreciate the support of the National Science Foundation (#0352519). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We are grateful to Liz Alix, Pamela Yeagley, and Lois Phillips for their contributions. For more information on the project, refer to the project Website at http://items.education.ucsb.edu or contact the first author at rzwick@education.ucsb.edu.

Note

[1] These three organizations subsequently collaborated on the development of a manual, *Interpreting and Communicating Assessment Results* (National Council on Measurement in Education, 1997). The project was led by Barbara S. Plake and James C. Impara and was funded by a grant to NCME from the W. K. Kellogg Foundation.

References

Adams, J. E., & Copland, M. A. (2005). *When learning counts: Rethinking licenses for school leaders*. Available at http://www.ncsl.org/print/educ/WhenLearningCounts_Adams.pdf.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Federation of Teachers, National Council on Measurement in Education, & National Education Association (1990). *Standards for teacher competence in educational assessment of student*. Available at http://ericsys.uncg.edu/AFT-NEA-AERA-Standards for Teacher Competence in Education.htm.

Begleiter, M. (2001). *From word to image: Storyboarding and the filmmaking process*. Studio City, CA: Michael Weise Productions.

Boudett, K. P., City, E. A., & Murnane, R. J. (Eds.) (2005). *Data wise: A step-by-step guide to using assessment results to improve teaching and learning*. Cambridge, MA: Harvard Education Press.

Brown, T., & Daw, R. (2004). *How do K-12 teachers and administrators feel about standardized testing and how well can they interpret results of such tests?* Unpublished report, Santa Barbara, CA: University of California.

Chance, B. (2002). Components of statistical thinking and implications for instruction and assessment. *Journal of Statistics Education, 10* (3). Available at www.amstat.org/publications/jse/v10n3/chance.html.

Confrey, J., Makar, K., & Kazak, S. (2004). Undertaking data analysis of student outcomes as professional development for teachers. *International Reviews on Mathematical Education (Zentralblatt für Didaktik der Mathematik), 36*(1), 32–40.

delMas, R. C., Garfield, J., & Chance, B. L. (1999). A model of classroom research in action: Developing simulation activities to improve students' statistical reasoning. *Journal of Statistics Education, 7* (3). Available at http://www.amstat.org/publications/jse/secure/v7n3/delmas.cfm.

English, L. D. (1997). *Mathematical reasoning: Analogies, metaphors, and images*. Mahwah, NJ: Erlbaum.

Finzer, W. (2001). *Fathom! (Version 1.12)* [Computer Software]. Emeryville, CA: Key Curriculum Press.

Garfield, J., delMas, R. C., & Chance, B. L. (2003). *The Web-based ARTIST: Assessment resource tools for improving statistical thinking*. Presented at the annual meeting of the American Educational Research Association, Chicago. Available at http://www.gen.umn.edu/artist/publications.html.

Hammerman, J., & Rubin, A. (2002). Visualizing a statistical world. *Hands On! 25* (2), 1–23. Available at http://www.terc.edu/downloads/Ho_Fall_o2.pdf.

Hammerman, J., & Rubin, A. (2004). Strategies for managing statistical complexity with new software tools. *Statistics Education Research Journal, 3*(2), 17–41.

Impara, J. C. (1993). *Joint Committee on Competency Standards in Student Assessment for Educational Administrators update: Assessment survey results*. Paper presented at the annual meeting of the National Council in Educational Measurement. Available at www.eric.ed.gov.

Jennings, J. (2002). New leadership for new standards. *Leaders Count Report*, Wallace Readers Digest Funds, Spring/Summer issue. Available at http://www.ctredpol.org.

Konold, C., & Miller, C. (1994). *Prob Sim®: A probability simulation program* [computer software]. Santa Barbara, CA: Intellimation Library for the Macintosh.

Konold, C., & Miller, C. (2004). *Tinkerplots (version 1.0)* [computer software]. Emeryville, CA: Key Curriculum Press.

Lukin, L. E., Bandalos, D. L., Eckhout, T. J., & Mickelson, K. (2004). Facilitating the development of assessment literacy. *Educational Measurement: Issues and Practice, 23*(2), 26–32.

Lundeberg, M. A., Levin, B. B., & Harrington, H. L. (Eds.) (1999). *Who learns what from cases and how? The research base for teaching and learning with cases*. Mahwah, NJ: Erlbaum.

Makar, K., & Confrey, J. (2002). *Comparing two distributions: Investigating secondary teachers' statistical thinking*. Presented at the Sixth International Conference on Teaching Statistics, Cape Town, South Africa. Available at http://www.stat.auckland.ac.nz/~iase/publications/1/10_18_ma.pdf.

Mayer, R. E. (2001). *Multimedia learning.* Cambridge, UK: Cambridge University Press.

Mayer, R. E. (2003). The promise of multimedia learning: Using the same instructional design methods across different media. *Learning and Instruction*, 13, 125–139.

Mills, J. (2002). Using computer simulation methods to teach statistics: A review of the literature. *Journal of Statistics Education, 10*(1). Available at http://www.amstat.org/publications/jse/v10n1/mills.html.

Moreno, R., Mayer, R., Spires, H., & Lester, J. (2001). The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents? *Cognition and Instruction, 19*, 177–213.

Narayanan, N. H., & Hegarty, M. (2002). Multimedia design for communication of dynamic information. *International Journal of Human-Computer Studies, 57,* 279–315.

National Council on Measurement in Education (1997). *Interpreting and communicating assessment results: Professional development resource materials*. Washington, DC: Author.

National Council on Measurement in Education Ad Hoc Committee on the Development of a Code of Ethics (1995). *Code of professional responsibilities in educational measurement*. Available at www.ncme.org.

Owen, J. M. (2007). *Program evaluation: Forms and approaches* (3rd ed.) New York: Guilford.

Pedulla, J., Abrams, L., Madaus, G., Russell, M., Ramos, M., & Miao, J. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers*. Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

Popham, W. J. (2006a). *Assessment for educational leaders*. Boston: Pearson.

Popham, W. J. (2006b). *Mastering Assessment: A self-service system for educators*. New York: Routledge.

Snell, J. L., & Finn, J. (1992). A course called "Chance." *Chance Magazine, 5*(3–4), 12–16.

Stiggins, R. (2002). Assessment for learning. *Education Week, 21*(26), 30, 32–33.

Stiggins, R., & Chappuis, J. (2005). Using student-involved classroom assessment to close achievement gaps. *Theory Into Practice, 44*(1), 11–18.

Wayman, J. C., Stringfield, S., & Yakimowski, M. (2004). *Software enabling school improvement through the analysis of student data* (Report No. 67). Baltimore, MD: Johns Hopkins University, Center for Research on the Education of Students Placed At Risk. Available at http://www.csos.jhu.edu/crespar/techReports/Report67.pdf.