Published in final edited form as:

Prev Sci. 2010 September ; 11(3): 275–286. doi:10.1007/s11121-010-0174-5.

Studying Implementation Quality of a School-Based Prevention Curriculum in Frontier Alaska: Application of Video-Recorded Observations and Expert Panel Judgment

Knowlton W. Johnson,

Alaska Office, Pacific Institute for Research & Evaluation, 4111 Minnesota Drive, Anchorage, AK 99503, USA

Kristen A. Ogilvie,

Alaska Office, Pacific Institute for Research & Evaluation, 4111 Minnesota Drive, Anchorage, AK 99503, USA

David A. Collins,

Louisville Center, Pacific Institute for Research & Evaluation, 1300 South 4th Street, Suite 300, Louisville, KY 40208, USA

Stephen R. Shamblen,

Louisville Center, Pacific Institute for Research & Evaluation, 1300 South 4th Street, Suite 300, Louisville, KY 40208, USA

Lisa G. Dirks,

Alaska Office, Pacific Institute for Research & Evaluation, 4111 Minnesota Drive, Anchorage, AK 99503, USA

Christopher L. Ringwalt, and

Chapel Hill Center, Pacific Institute for Research & Evaluation, 1516 East Franklin Street, Suite 200, Chapel Hill, NC 27514, USA

Jennifer J. Norland

Alaska Office, Pacific Institute for Research & Evaluation, 4111 Minnesota Drive, Anchorage, AK 99503, USA

Knowlton W. Johnson: kwjohnson@pire.org

Abstract

This study assesses the implementation quality of *Think Smart*, a school-based drug prevention curriculum that was designed to reduce use of harmful legal products (HLPs; e.g., inhalants and over-the-counter drugs), alcohol, tobacco, and other drugs among 5th- and 6th-grade students in frontier Alaska. Participating in the study were eight communities that took part in a larger randomized control trial to assess the short-term effects of the *Think Smart* curriculum. Video-recorded observations of the 12 core and 3 booster lessons were conducted in 20 classrooms. Ninety-five sessions were randomly selected from 228 usable videodiscs, and two pairs of researchers observed each video recording to code level of dosage, adherence to curriculum design, and teachers' delivery skills. Inter-rater reliability for all implementation quality measures was very high. An expert panel consisting of 16 scientists reviewed the results of the implementation study and concluded that the level of dosage and adherence to the curriculum design was at least as high as those yielded by similar studies. However, the panel assessed the

delivery quality to be only marginal in comparison to results of other studies. The experts concluded that the implementation quality of the *Think Smart* curriculum was adequate even though the teachers' delivery skills were only marginal. A bootstrapping analysis, in which 1,000 samples were drawn for each implementation quality result, found the expert judgments to be reliable. The authors conclude that despite some limitations, video-recorded observations, as well as expert judgment, provide strong methodologies that should be considered for future implementation quality studies.

Keywords

Implementation quality; Video recorded observations; Expert panel assessment; Youths' harmful legal product use; School-based prevention

Introduction

There is a growing emphasis in the prevention field on both the use of research-based prevention curricula in school settings and on implementing these curricula with high quality. Since 2002, the Safe and Drug-Free Schools and Communities Act has required that its funds be used for programs proven to be effective (NCLB 2004). There are now several federal registries that identify prevention programs that have been evaluated with methodological rigor and have yielded positive behavioral effects (e.g., CSAP 2001; OJJDP 2002; Robertson et al. 2003; U.S. DOE 2001). A number of authors have also discussed the importance of delivering prevention programs in the way in which they are intended (Dane and Schneider 1998; Domitrovich and Greenberg 2000; Dusenbury et al. 2003; McGrew et al. 1994). In addition, studies and reviews (Backer 2001; Berman et al. 1975; Fullan and Pomfret 1977; McCormick et al. 1995), as well as guides to effective school-based drug prevention curricula (Bosworth 2000; Centers for Disease Control and Prevention 1994; Drug Strategies 1999; Robertson et al. 2003), strongly emphasize the importance of implementation quality.

In contrast, reviews of the research on fidelity and implementation quality (Dusenbury et al. 2003; Dusenbury et al. 2005; Lillehoj et al. 2004; Ringwalt et al. 2003) have consistently found that school-based prevention curricula tend to be implemented with great variability in real-world settings, and that teachers' adaptations may translate into less desirable implementation quality. Some argue that a few modifications are not only inevitable, but may be desirable to meet the differing needs of implementers, participants, or their social, cultural, or environmental contexts (Berman and McLaughlin 1976; Domitrovich and Greenberg 2000; Ridgely and Jerrell 1996). Supporting this view, some studies suggest that lack of fidelity may not always threaten the integrity of an intervention (Dane and Schneider 1998; McGrew et al. 1994) and that some degree of curricular adaptation is indeed normative (Botvin et al. 2000; Ringwalt et al. 2003; Spoth et al. 1988; Tricker and Davis 1988). Nevertheless, researchers are focusing more on understanding how programs are implemented, and national registries of effective programs are beginning to require that implementation be carefully assessed, reported, and integrated into discussions of program outcomes (e.g., NREPP 2007).

This article presents results that contribute to a growing body of literature on implementation quality in the use of school-based prevention curricula. Two innovative methods of studying implementation quality are highlighted: video-recorded observations and an expert panel assessment. The research setting for this study comprised frontier Alaska schools in remote communities that largely have no road access to urban centers of the state. The analysis used data from a randomized control trial funded by the National

Institute on Drug Abuse (NIDA) to study the implementation and effects of a school prevention curriculum (*Think Smart*) on pre-adolescents' use of inhalants and other legal products that adolescents in frontier Alaska use to get high. The outcome results have been reported by Johnson and his colleagues (2009). This study in frontier Alaska communities is especially relevant for schools in rural and remote areas of the world.

The specific research questions that address the implementation of the *Think Smart* curriculum and have guided this present study are:

1. How reliable are data produced by video-recorded observations as a means to assess the implementation quality of the *Think Smart* curriculum?
2. What is the level of implementation quality of this curriculum as measured by dosage, adherence, and delivery skills?
3. How does the level and adequacy of implementation quality of this curriculum compare to the quality of implementation as reported by experts' judgments of studies of similar prevention curricula?

Implementation Quality Research in Perspective

Both implementation quality and fidelity are discussed extensively in the literature, and there is a great deal of overlap between the two concepts. Yeaton and Sechrest (1981) defined fidelity as "the degree to which treatment is delivered as intended" (p. 160). Dane and Schneider (1998) discussed elements of implementation fidelity in terms of a typology that includes adherence, dosage, quality of program delivery, participant responsiveness, and program differentiation (whether key features are presented that distinguish the program from others). Domitrovich and Greenberg (2000), in a mental health study that targeted school-age children, shifted their focus from implementation fidelity as an overarching construct to implementation quality. Later, Dusenbury and her colleagues (2005) suggested that the following key variables are likely to be important to the achievement of high implementation quality: dosage (provision of sufficient exposure), adherence (completing content and delivery as designed), quality of the process (actively engaging participants), and adaptations (modifications designed to meet developmental and cultural needs). For the present study, we define implementation quality in terms of (1) dosage, (2) adherence, and (3) delivery quality. This operationalization of implementation quality was based on the researchers' decision to include the three most dominant measures that appear in the literature, and to use multiple indicators to enhance the precision of measurement.

Collecting Implementation Data—There have been multiple approaches to the collection of implementation quality data. Hansen and McNeal (1999) noted that most studies have relied almost exclusively on teacher and administrator self-reports. Dusenbury and colleagues (2003) reported that some of the more traditional sources of data concerning fidelity include direct observations, teachers' self-reports, and students' reports of teachers' implementation. They conclude that while data from teachers appear to be more comprehensive, data from trained observers tend to have higher validity. In this regard, Lillehoj et al. (2004) suggest that program implementers tend to overestimate the fidelity with which they have implemented their programs, when their ratings are compared to those of independent observers.

Dane and Schneider (1998) have suggested that there is a stronger link between implementation assessment and outcomes when trained observers are the data source, and they attributed this result to the effects of social desirability bias on program providers' self-reports. Lillehoj et al. (2004) found that providers tended to report higher levels of implementation fidelity than did independent observers. In addition, self-reported

implementation scores had limited variance, which restricted the investigators' ability to detect relationships between implementation and outcomes in the population served (Lillehoj et al. 2004). Melde et al. (2006) described a process evaluation of a school-based victimization prevention program that utilized provider surveys and observations. They reported that, in general, the data derived from both data collection techniques were similar.

Although these approaches to data collection remain the standard in implementation quality assessment, their problems suggest that alternative strategies might prove useful to improve data quality. One alternative is the use of video recordings in place of direct observations. To date, there has been little discussion on the use of this method for assessing the fidelity of programs implemented in classroom settings. Sobol and colleagues (1989) rated instructors leading students in creating role-plays, which were video-recorded and then coded by five raters. Instructors were rated on two components of integrity: the degree to which they followed curriculum guidelines for role-plays and the quality of their delivery. The investigators found that those teachers who taught with higher integrity were more animated, articulate, and confident. The investigators also demonstrated the feasibility of obtaining reliable measures of implementation quality using recorded observations, suggesting this was a methodological advance over previous studies that relied primarily on self-report. More recently, Pankratz and her colleagues (2006), Ringwalt et al. (2009), and Giles et al. (2008) used video-recorded lessons to measure adherence and adaptation of an underage alcohol use prevention curriculum. Teachers were provided with cameras and video discs to record their lessons. Inter-rater reliability as measured by Cohen's kappa was high for adherence and adaptations to teaching methods, but reliability was only moderate for adaptations to lesson content.

Video observation provides both comprehensive and reliable data in diverse settings. Video recording for the purposes of studying implementation quality has a very practical function in situations where live observers are unavailable for a variety of reasons. Researchers in rural and frontier settings, for instance, are challenged to find, train, and sustain qualified research personnel (Bierman et al. 1997). Local staff turnover and the resulting need to recruit and train new personnel can delay project progress and significantly increase costs. The use of video-recording emerged as a possibly viable data collection method for frontier Alaska school systems.

Assessing the Adequacy of Implementation Quality

There has been little discussion regarding how to assess the adequacy of implementation quality. For instance, Spoth and colleagues (2007) assessed the level of implementation quality using only the judgments of the research team, which described implementation quality with only the nonspecific terms "high" versus "low." Others have linked implementation quality to program outcomes, implying that there is a positive relationship between the two, so that desirable outcomes indicate adequate implementation. Melde and colleagues (2006) analyzed observation reports by school, classifying their schools along a continuum of high, medium, and low degrees of fidelity. The researchers specified *a priori* that a high degree of fidelity indicated that all or nearly all of the lesson information was delivered within the time recommended and with effective classroom management; medium fidelity indicated schools in which time or classroom management issues interfered with proper implementation; and low fidelity indicated those schools in which the program was not implemented as intended.

The use of expert panels to study implementation quality constitutes an emerging assessment methodology. In their seminal work on the subject, Meyer and Booker (2001) describe eliciting expert judgment as a fundamental methodology in addressing "technical questions" that for any one of a number of reasons cannot be answered with experimental or

observational data. These authors cite examples of the use of expert panels in the physical sciences—for example, to review the frequency and magnitude of earthquakes that are based on a review of seismic-tectonic information and then make judgments about the accuracy of these estimates based on their seismic research experience.

There has been limited use of expert panel assessments in the social and behavioral sciences, although several exceptions were found in the pertinent literature. Johnson (1990) used an expert panel drawn from the membership of the American Evaluation Association to assess whether the decision makers of 440 criminal justice agencies adequately used research results in their decision-making processes. Brandon and colleagues (2008) used an expert panel to evaluate the quality of implementation of a science curriculum in middle schools. Gingiss et al (2006) asked an expert panel to participate in defining questions, response options, and scoring procedures for a *Bridge-It's* mathematical model that are designed to produce a quantitative statement of probability of future implementation fidelity of school-based prevention programs.

For the present study, we utilized an expert panel of social and behavioral scientists, most of whom were prevention researchers, to make judgments about the level and adequacy of implementation quality of the *Think Smart* prevention curriculum, in comparison to their own research and that of others.

Methods

Community and School Participants

The study engaged eight communities in Alaska whose elementary schools implemented the *Think Smart* substance use prevention curriculum in the 2006–2007 school year. The communities that participated in the *Think Smart* study represent part of the 52% of Alaska's population that is identified as frontier (National Center for Frontier Communities 2007). All but one of the study communities are located off the road system (i.e., air-or water-craft are the only means of access). The communities range in population from 400 to 4,300. Four communities have predominately Alaska Native populations, including three villages with Alaska Native populations greater than 90%. Three communities comprise a majority White population and one community has a 48% Asian/Pacific Islander population. Seven school districts serve the participating eight communities; one school district serves two communities.

Within the eight communities, 19 teachers were responsible for delivering the *Think Smart* curriculum, of whom one taught two classes for a total of 20. However, one teacher did not provide usable data. Most teachers had completed at least some graduate-level education (63%). Sixteen were White (84%) and the others were Alaska Native or American Indian; all were non-Hispanic. Teachers were predominately female (53%), of whom most were over the age of 36 (74%). Six teachers had lived in the community for less than a year, five had lived in the community from 1–10 years, and seven had lived in the community for over 10 years. Almost half of the teachers (47%) had over 15 years of experience in the field of education while 32% had less than 1–5 years of experience. Forty-seven percent of the teachers reported having some experience implementing school-based prevention programs.

Think Smart Curriculum

The *Think Smart* curriculum constitutes a modified form of the Personal Intervention Curriculum, a research-based drug prevention curriculum developed by Steven Schinke for a Pacific Northwest American Indian population (Ogilvie et al. 2006; Schinke et al. 1988, 2000). The curriculum, in turn, is based on Botvins' Life Skills Training. Study team members involved in the adaptation included scientists from the research team, Sociometrics

(the firm that holds the rights to Schinke's curriculum), as well as Alaska Native and other Alaska consultants. The curriculum was not adapted for Alaska Natives specifically, but for frontier Alaska more generally.

In making the adaptations, the study team utilized Ringwalt and Bliss's (2006) synthesis of three types of adaptations: surface, deep, and evidential. To make the curriculum more relevant to Alaska, surface adaptations included Alaska-specific visuals and examples; deep adaptations involved integrating the values of Alaska people (e.g., excluding "BINGO," which was in the original curriculum, because gambling is a serious problem in many rural Alaskan communities), and evidential adaptations provided more Alaska-specific statistics. Lessons learned in a prior NIDA-funded feasibility study that included an earlier version of *Think Smart* were incorporated into a revised curriculum that was implemented in this study (Johnson et al. 2007).

The *Think Smart* curriculum consists of 12 core lessons, including introductory and concluding lessons, and 3 booster lessons implemented 2 to 3 months later. The core lessons include lessons on stereotypes and drug facts and the introduction of a problem-solving model known as S.O.D.A.S. (Stop, Options, Decide, Act, Self-Talk), which emphasizes refusal and self-assertiveness skills. The final core lessons and the booster lessons involve practicing this problem-solving model. The curriculum involves many interactive activities, including role-plays and other engaging methods.

In this study, the classroom teachers taught the core and booster lessons weekly as 1-hour lessons in 5th- and 6th-grade classrooms. The core lessons were taught between October 2006 and February 2007 in the eight communities, and the booster lessons were taught 2 to 3 months later during March through May 2007. The teachers received a 2-day, on-site training from the intervention agency partner who, together with a trained, local Prevention Project Coordinator (PPC), offered technical assistance to the teachers throughout the curriculum's implementation. A complete teaching kit with all the items necessary to implement the interactive curriculum (e.g., laminated role play cards, timers) was provided to each teacher. Teachers were also given video cameras to record all of the lessons in order to facilitate the study of implementation quality.

Implementation Quality Measures

Dosage, adherence, and delivery quality indicators were adapted from the standards developed by Pankratz and her colleagues (2006). These standards included coding one indicator of dosage, four indicators for adherence, and six indicators for delivery quality. Table 1 presents the implementation quality measures used in this study by indicators, and operational definitions of each indicator.

Dosage was measured only if the instructor attempted, or initiated, any activity in the 13 lessons that were observed. *Adherence* was coded at the step level and aggregated to the activity level if more than one step was involved in an activity. The operational definitions for adherence (presented in Table 1) include four single indicators and a composite index that specifies the proportion of changes, additions, omissions, and new delivery methods.

Delivery quality assessment was adapted from the standards used by Giles and his colleagues (2008) that enumerated the number of skills displayed per activity in each classroom. These standards seek to document interactive teaching skills. Following Giles (2008), ten original categories were coded, of which the six categories specified in Table 1 were found to be statistically significant in an exploratory factor analysis to be described later.

Data and Sample

Implementation Quality—Teachers video recorded each *Think Smart* lesson taught, excluding Lessons 1 and 12, which constituted introductory and congratulatory lessons. A local project data coordinator collected the videodiscs and mailed them to the research team in the project office in Anchorage. Research staff examined each videodisc to determine the quality of the recording. Some discs were of poor quality and were not included in the observation sample, and others were not received in the mail. The length of the video-recordings was dependent on the content and interactivity of the lesson and ranged from 30 minutes to just over one hour. Of a possible 260 total videodiscs, 228 (or 88%) were of usable quality. While the usable discs were evenly distributed across the lessons, some teachers did not provide as many discs as others. Of 20 possible discs, each lesson yielded between 16 and 19 usable discs with an average of 16.9. Of the 13 possible discs, each teacher mailed between 0 and 13 usable discs with an average of 11 discs. Twelve teachers provided usable discs for all 13 lessons, and only three provided fewer than 10 usable discs.

We randomly chose 95 of the 228 sessions as a sample for the implementation quality study. The sample represented the core lessons (2 through 11) of the curriculum as well as the three booster lessons administered three months later. Ninety-five sessions were chosen, in order to achieve an acceptable margin of error for producing accurate estimates. That is, assuming a 95% margin of error, a finite population of 228 sessions, and the most conservative base rate of a the variance of any event (50%), our sample yielded at most a 7.72% margin of error for our estimates. These procedures led to an adequate sample, as on average there were 7.31 (SD=2.10, Mode=6) observations for each lesson and 5.00 (SD=2.05, Mode=5) observations for each teacher.

Two pairs of coders evaluated each video-recorded observation of the random sample. One pair observed dosage and fidelity while the other observed teacher delivery skills. All coders received interactive training on using the coding instruments. In order to establish that coders were interpreting the curriculum in similar ways, they initially rated a small sample of sessions and then compared their responses. From these sessions, a set of decision rules was developed that guided the coding process. To further ensure consistency, coders met periodically to establish consensus during the months in which the coding process took place to reinforce the established decision rules. Each pair of coders established consensus on an average of 19 sessions.

Expert Panel—The assessment of the *Think Smart* implementation quality study entailed the use of expert judgments. Meyer and Booker (2001) defines an expert judgment as data given by an expert in response to a technical problem. The expert panel for this study consisted of scientists who had conducted research focusing on implementation quality or fidelity and who indicated they were knowledgeable about the topic matter. The process by which we identified the expert panel is consistent with Nevo's (1989) and Meyer and Booker's (2001) criteria for forming expert groups (e.g., membership should be diverse, and should have as wide a range of expertise as appropriate and possible). Sixteen scientists agreed to serve on the expert panel and, for participation, \$100 was offered to the charity of their choice. The 16 expert panel respondents all had doctoral degrees. The majority of the respondents were female (63%) and worked in university settings (69%). The participants averaged 15–20 years of research experience. Further, on average, the respondents were involved with 10.81 (SD=14.72) curriculum development studies and 6.00 (SD=3.74) implementation development studies. The panel members were not familiar with Alaska culture other than what we provided in our handout packets. Further, the panel was 100% non-Hispanic White. We believe none of these characteristics biased their judgments

because the assessment was supposed to be objective and based solely on their implementation quality research experience.

We implemented the following preliminary steps before we began collecting data. We began by sending each panel member a packet of material describing the project, expert panel tasks, and the implementation quality results (in both table and graph formats) that summarized the observers' assessments of the video-recorded classroom sessions. The tables and figures included information on the observers' assessments of the level of dosage and adherence to the curriculum design, as well as ratings of delivery skills during curriculum implementation. We also sent a Power-Point presentation for the panel members to refer to during the telephone interview. The presentation included an abbreviated version of the information in the packet as well as the assessment questions.

Approximately 1 to 2 weeks after sending the materials, we conducted a telephone interview with each of the expert panel members to obtain their ratings of the adequacy of implementation quality of the *Think Smart* curriculum. The experts assessed the level and adequacy of implementation quality for dosage, adherence, and delivery skills, using as a benchmark their own research or that of others involved in implementation research. The instructions to the experts called for them to rate levels (of dosage, adherence and delivery skills) on a five-point scale that ranged from "much lower than average" to "much higher than average," and to rate adequacy on a four-point scale (inadequate, marginally adequate, moderately adequate, and adequate). Each expert also assessed overall adequacy of quality ratings for the entire curriculum implementation considered in aggregate. Once this step was completed, the success ratings provided by the panel were compiled for analysis and presentation.

Results

Inter-Rater Reliability of the Implementation Quality Data

The first research question concerns an assessment of the reliability of data used in the implementation quality study. Table 2 presents the inter-rater reliability results that were based on the 95 sessions, excluding those used to establish coder consistency. Since the dosage and adherence indicators were nominal level measures, we used percentage agreement and Cohen's kappa as a test of strength of association (Cohen 1960; Fleiss 1971). The five dosage and adherence indicators ranged from 96% to 99% agreement, and the inter-rater kappa ranged from 0.88 to 0.97. Pearson's *r* and intra class correlations were used to determine the inter-rater reliability of delivery skills, which are continuous indicators (Fleiss 1971; Shrout and Fleiss 1979). Measures of the inter-rater reliability of the six delivery quality indicators were equally high, with Pearson correlations ranging from 0.89 to 0.99 and intra class correlations ranging from 0.87 to 0.99.

Level of Implementation Quality

The second research question focused on an assessment of the levels of dosage, adherence, and delivery skills used to implement the *Think Smart* curriculum.

Dosage and Adherence—For levels of dosage and adherence to the curriculum design, activities across topic and lesson-level indicators were aggregated and computed as percentages. For dosage, we aggregated teachers' attempts of an activity within each observed session and computed a percentage by dividing the total possible attempts by actual attempts; the resulting ratio constituted the average level of dosage per session. The session-level content adherence indicators began as ratings of 'changes,' 'omissions,' 'additions,' and 'new methods' generated by two coders watching

recordings of teachers implementing the lesson activities. Activity- and session-level dosage and fidelity indicators were computed as percentages. We subtracted the session-level indicators from 1 to measure positive implementation quality for dosage, adherence, and delivery quality. For adherence, we created a composite outcome score, which constitutes an overall measure of adherence relating to the percentage of observations with no changes, omissions, additions, or any new methods. The final average dosage and adherence results were divided into quintiles for ease of presentation.

Table 3 presents the level of dosage and adherence to the *Think Smart* curriculum design. In this table, the quintiles of the percent of activities across sessions can be found in Column 1, followed by the percentage of observations for each of the dosage and adherence indicators, including the composite adherence index in Columns 2–7. The mean percentage for each indicator is also presented. For dosage, the mean percentage was 82.2%; 72.6% of those who attempted or completed a lesson fell within the 80–100% quintile. There was a noticeable variation in the level of adherence to the curriculum design. The percentages of activities in which no new methods (87.6%) or content changes (82%) occurred were higher than no additions (76%) and no omissions (59.3%). However, the composite score, which takes into consideration all four adherence indicators, shows that adherence to the curriculum was only marginal. It should be noted that these measures do not take into consideration the reason for non-adherence to any particular activity.

Delivery Skills—Delivery skill indicators comprised counts of observed occurrences in the recorded activity delivery. Initially, ten skills were assessed. A multi-level exploratory factor analysis using LISREL was conducted, in which lesson-level data were nested in teachers. We found that six of nine indicators formed an unidimensional scale with factor loadings of .50 and higher and an alpha reliability of 0.75. The data fit was good when the categories of enthusiastic and less enthusiastic praise were combined into a single indicator—Chi square=13.6 ($p=0.75$), RMSEA=0.00. The skills used in the following analysis include: Praise, Accepts Ideas of Students, Uses Ideas of Students, Asks Original Question, Asks Repeated Question, Asks Probing Question, and a mean composite comprised of the six individual measures. (See Table 1 for examples of these skills.) Z-scores were computed from raw observed occurrences as a method of identifying outlying observations. For each delivery skill, any observations above three standard deviations from the mean were Winsorized; no more than 2 ratings for the 95 observations per indicator met this criterion. The observations were standardized on a scale of 0–100 and then organized into quintiles. The three remaining skill indicators that measured self-disclosure of personal anecdotes or correcting students' behaviors did not form a second factor that was sufficiently valid and reliable to retain.

Table 4 shows the average number of occurrences (and the standard deviation), as well as the range (minimum and maximum). All of the results presented are based on the mean of the average delivery skills across observed sessions. The highest number of occurrences observed per session was Asked Original Questions (28) and Accepted Ideas of Students (24). Teachers asked repeated questions per session (7), gave praise (4), asked probing questions (3), and used ideas of students (2) much less frequently. There was considerable variation across classroom teachers in this regard; the most variation was in acceptance of ideas of students ($SD=10.4$) and the least variation was in the use of students' ideas ($SD=1.5$).

Expert Panel Assessment

The third research question concerns whether expert judgments about teachers' level and adequacy of implementation quality differ from *a priori* expectations of the research team.

These expectations were set at an average rating of 3 for level of quality per measure (0=Much lower than average, 2=Average, and 4=Much higher than average) and an average rating of 2.5 for adequacy (0=Inadequate, 1=Marginally Adequate, 2= Moderately Adequate, and 3=Adequate). It was assumed that a higher than average level of quality (i.e., 3 on a 0–4 scale) and adequacy (2.5 on a 0–3 scale) across the 16 experts should produce the desired change in the targeted outcomes. Single sample *t*-tests were performed comparing observed to expected values. For this statistical analysis, we tested whether the observed-average expert rating differed significantly from the expected rating.

Expert panels usually consist of a small number of members, which precludes performing analyses from which inferences can be drawn (Meyer and Booker 2001). To increase our confidence in the results from our expert panel of 16 members, we analyzed the observed data and then performed a bootstrap analysis to improve confidence in our results (Meyer and Booker 2001). Using Excel, we drew 1,000 bootstrapped samples of size 16, sampling with replacement, for each of our results. We calculated average test values across all bootstrapped samples, except for *p* values, which we determined from the average *t*-statistic.

Table 5 presents the expert panel assessment of both the level and adequacy of the level of implementation quality of the *Think Smart* curriculum. For the level of implementation quality, the 16 experts rated the level of dosage, adherence, and delivery skills of the *Think Smart* curriculum. These results were replicated in the bootstrapping analysis described earlier. The average expert's dosage and adherence ratings (four combined scores) for the level of quality in comparison to other studies was higher than average (3), which also constituted the expected score for both measures. When we bootstrapped these results for the 1,000 sets of 16 sampled, the results were essentially the same; only 4% of the sample of dosage results was significantly different when comparing actual with expected, and adherence results were significantly different for 5% of the 1,000 sample. Contrary to our expectations, the panel rated the level of teachers' use of delivery skills in the 19 classrooms lower than expected (observed=2.47 vs. expected=3). While this rating was still above average, the mean difference of $-.53$ was statistically different in both the actual and bootstrap analyses ($t=3.44$; 3.60 respectively). The Cohen *d* of $-.89$ and $-.93$ showed this to be a large difference. We also found that this difference was consistent for 98% of the 1,000 samples in the bootstrapping analysis.

For the adequacy of the level of the implementation quality of the *Think Smart* curriculum, the mean ratings for dosage, adherence, delivery skills, and overall curriculum implementation met or exceeded the expectations set by the research team, even though the panel of experts rated the level of use of delivery skills lower than we expected. The test of the small difference for the four measures was not statistically significant for either the actual or bootstrapped results.

Discussion

This study investigates the use of ratings of video-recorded observations by trained coders and judgments by an expert panel to determine the level and adequacy of implementation quality of a school-based prevention curriculum (*Think Smart*), which was designed to prevent substance use among pre-adolescents in frontier Alaska. The research questions and corresponding results focused on the quality of implementation data, actual implementation quality, and expert judgments of the level and adequacy of implementation. We found the inter-rater reliability of dosage, adherence, and delivery skills to be excellent, which means that the reliability estimates exceeded the expectations of the investigators. We also found that the expert panel judged the level of implementation quality as acceptable, and that the level of dosage and adherence met those of most published studies of fidelity and

implementation quality. However, the expert panel found the quality of curriculum delivery skills did not meet their expectations. Nevertheless, the experts found that the adequacy of implementation, which included dosage, adherence, and delivery skills, was acceptable. When the experts' judgments of the *Think Smart* curriculum's implementation quality were compared in a statistical analysis with standards set by the research team, most expectations were met. Further, when the expert panel results and the comparisons with the research team's expectations were replicated across 1,000 samples by means of a bootstrapping analysis, these results showed that the expert panel results were reliable.

One of this study's notable strengths is that sessions in which the *Think Smart* curriculum was implemented were video recorded instead of relying teacher self-reports or direct observation. To date, direct observation to collect fidelity or implementation quality data does not constitute the dominant data collection method, even though such observations are thought to be more objective, valid, and reliable than implementers' self-reports (Dusenbury et al. 2003; Rohrbach et al. 2007). However, the discussion is now shifting from direct (that is, in person) to video-recorded observations (Brandon et al. 2008; Pankratz et al. 2006). Although video recording, as a method of observation, can serve as a useful mechanism for examining implementation quality, a variety of drawbacks are apparent.

In our study, despite training on placement and operation of the video cameras, considerable variation existed in both. These discrepancies resulted in recordings that varied in both sound quality and visual clarity. The curriculum includes topics that require the instructor to post materials on the board or state aloud certain instructions. While a majority of the recordings had clear audiovisual distinction, it was sometimes difficult to discern whether or not an instructor implemented steps that were designed to be quoted verbatim from the curriculum guide. Another drawback related to the instructor's control of the camera. In several classrooms, instructors shut off their cameras while still teaching the lesson. In some cases, the cameras were turned on in the midst of an activity or sub-activity. Some instructors may have delivered parts of the curriculum while the camera was off, leaving coders to presume that they did not complete those segments of the curriculum. If this occurred, it would have inadvertently decreased the observed total percentage completed for some of the sessions. These problems highlight the issue that even under ideal conditions, video recording may omit some aspects of instruction that might be captured by direct observation. There were additional logistical challenges associated with video recording. Teachers in 8 of the 260 possible sessions experienced technical difficulties (e.g., the camera did not record) that precluded the use of those sessions in the sample. An additional number of sessions were not recorded in their entirety, primarily because the teacher exceeded the 60-minute length of the videodisc. Finally, teachers in three classrooms did not submit all of their videodiscs, despite multiple efforts by the study team to collect them.

Despite these problems, 88% of the videodiscs were deemed acceptable. This study demonstrates that video-recording has a number of clear benefits. Direct observation in 20 classrooms would have been impossible given the distance and cost of such a methodology in rural Alaska, as we discovered in a previous effort to recruit, train, and retain observers in another study in four Alaska communities. Another benefit was the ability to review, as needed, sections of the recordings for clarity and to comb through the video recordings as often as needed to resolve discrepancies between coders. Thus, the study was able to establish high inter-rater reliability and avoid coder drift. Despite the nuances of instructions that were lost in recording, the two video observers were able to code details for both fidelity and delivery, which would have presented a significant challenge for direct observers unless two had been present in the classroom. Finally, the video cameras were less intrusive for both teachers and students than direct observers would have been, which increased the likelihood that we captured authentic classroom implementation. For these

reasons, video-recorded observations represent a potentially valuable method to assess implementation quality in studies of both program effectiveness and diffusion.

A second strength of this study is that the observation results can be generalized to all curriculum lessons and all classrooms in which the *Think Smart* curriculum was implemented, because we drew a random sample of 95 sessions from a usable 228 sessions, which constituted 88% of the total of 260 sessions administered. Other observational studies have typically used a purposive sampling method that is based on varying criteria that limit generalization to either all lessons being taught or all classrooms in which the curriculum was implemented (e.g., Brandon et al. 2008; Dusenbury et al. 2005; Pankratz et al. 2006; Rohrbach et al. 2007). It should be noted, however, that the high cost and labor-intensive nature of the observations required to code video-recording sessions may limit the size of the sample selected in future studies. That would, in turn, reduce the investigators' ability to generalize from sample results. In our study the sampling error was plus or minus 7%, which was even higher than desired.

A third study strength is that the reliabilities found for nominal and continuous implementation quality measures were very high, ranging from 96–99% agreement, as assessed by Kappas (0.88–0.97), Pearson correlations (0.89–0.99) and intra class correlations (0.87–0.99). These inter-rater reliabilities exceeded direct and video-recorded observation published in the literature to date, although other studies have reported acceptable inter-rater reliabilities (Lillehoj et al. 2004; Pankratz et al. 2006; Rohrbach et al. 2007). We believe our high inter-rater reliability is a function of our coding protocols. That is, a consensus method established inter-coder agreement before moving to independent reviews and sustained a process to address coder drift.

Fourth, the *Think Smart* study measures a high level of detail about dosage, adherence, delivery skills, and overall curriculum implementation quality. We assessed implementation quality at both the step and activity levels, whereas most studies have focused on adaptations at the level of the entire lesson. (One notable exception in the prevention literature is the related work of Pankratz et al. (2006); Ringwalt et al. (2009); and Giles et al. (2008), whose video recording and coding methodology served as a model for our research.) Video recording allowed us to move back and forth in our coding to capture implementation detail at the step and activity levels for all substantive lessons. A further distinction is that our adherence measure, which comprised four indicators, more precisely measured adherence as designed, in that we assessed the degree to which there were no changes, omissions, additions, or new methods. Despite these advances, one study weakness was our inability to speak to the reasons why teachers made adaptations. While we concentrated on the coding of implementation quality, we acknowledge that it would also have been desirable if we could have assessed why adherence varied.

A fifth and final strength is the use of an expert panel to assess the level and adequacy of implementation quality. This methodology is relatively new to the social and behavioral sciences and seldom used. The method we used of interviewing our study's experts by telephone after providing them with an information packet was clearly beneficial; most of the experts saw these provisions as adequate information that allowed them to participate in these assessments with ease. Although we believe the use of an expert panel assessment provides value-added importance to the prevention field, several limitations should be noted. One is that whereas the panel members had published in the areas of curriculum implementation and fidelity, their experience may have been based on the use of different indicators and criteria than were used in the *Think Smart* study. For example, one expert noted being more familiar with, and therefore better able to assess, delivery quality in terms of indicators such as "participation by students encouraged by teacher or instructor."

Another noted that her expertise was more related to family-focused than school-based interventions, “where implementation quality tends to be more elaborated.” Some suggested more information on the context would have been helpful in their assessments; e.g., that support and leadership within the school administration could be key to actual implementation quality. To overcome these limitations a more interactive interview, or focus group, might have illuminated more in-depth insights into implementation quality. Some of these limitations were related to the fact that our experts did not have in-depth information on the curriculum itself. At the same time, the use of outside experts also meant that there was less likelihood of bias in their ratings of implementation quality than had they been associated with the curriculum.

Some of the experts also expressed having more difficulty in assessing the level and adequacy of delivery skills relative to dosage and adherence. For example, some noted that they would have benefited from having more information about what the curriculum emphasized and would like to have been provided with average numbers of occurrences of delivery skills. One expert also noted that he did not have any information about how the teachers included in this study stood in terms of their level of delivery skills in general (that is, outside of this curriculum). For future studies that involve an assessment of delivery skills with these measures, we recommend developing clear expectations for teachers and establishing a baseline standard to which to compare teacher skill.

In conclusion, the study of implementation quality of the *Think Smart* curriculum in frontier Alaska schools demonstrates that video-recorded observations provide a viable data collection method. Although there were some issues related to the quality of the videodiscs, lessons learned from this study should increase the quality of implementation data for studies that utilize video-recordings. While we also found that coding video-recorded observations could enhance the inter-rater reliability found in this study, this approach yields increased costs and problems associated with the intensiveness and iterative nature of the coding process. One possible remedy is to reduce the level of specificity of the implementation measures, thus reducing the resources required to rate videodiscs. Freeing up coding resources would also allow for increasing the sample size that, in turn, would improve the generalization of the implementation quality results. This study also demonstrates the feasibility of using an expert panel, especially in conjunction with a bootstrapping analysis, to increase the reliability of results of a small sample of experts. The lessons learned noted earlier should enhance future use of this implementation assessment method. A combination of video-recorded observations and expert panel assessment is especially appropriate for the study of school-based prevention curricula implementation in frontier communities. Finally, it is important to establish the predictive validity of implementation quality using observations of video recordings. Linking implementation quality to outcomes is the ultimate step in establishing the importance of implementation quality.

Acknowledgments

Research for and preparation of this manuscript were supported by the National Institute on Drug Abuse Grant #R01 DA019640, K. Johnson, PI. The authors wish to express their gratitude to Diane Ogilvie of Akeela, Inc. who collaborated on adapting the curriculum and managed its implementation during the study as well as the Alaska schools and their teachers, staff, students, and administrators that made this study possible. A debt of gratitude is also owed to the coders of curriculum. The authors finally wish to acknowledge the 16 expert panel members, some of whom also made valuable suggestions to improving an earlier version of the manuscript. These members are Drs. Paul Brandon, Linda Dusenbury, Abby Fagan, Steve Giles, Phyllis Gingiss, Mark Greenberg, Bill Hansen, Laura Hill, Anja Kurki, Torill Larsen, Catherine Lillehoj, Melinda Pankratz, Luanne Rohrbach, Richard Spoth, Steve Sussman, and Jennifer Unger.

References

- Backer, TE. Finding the balance: Program fidelity and adaptation in substance abuse prevention. Rockville, MD: Center for Substance Abuse Prevention; 2001.
- Berman, P.; Greenwood, PW.; McLaughlin, MW.; Pincus, J. Federal programs supporting educational change. Santa Monica CA: RAND; 1975.
- Berman P, McLaughlin MW. Implementation of educational innovation. *The Educational Forum*. 1976; 40:345–370.
- Bierman KL, Cole JD, Dodge KA, Greenberg MT, Lochman JE, McMahon RJ. Implementing a comprehensive program for the prevention of conduct problems in rural communities: The Fast Track experience. *American Journal of Community Psychology*. 1997; 25:493–514. [PubMed: 9338956]
- Bosworth, K. Protective schools: Linking drug abuse prevention with student success. A guide for educators, policy makers, and families. Tucson: University of Arizona; 2000.
- Botvin GJ, Griffin KW, Diaz T, Scheier LM, Williams C, Epstein JA. Preventing illicit drug use in adolescents: Long-term follow-up data from a randomized control trial of a school population. *Addictive Behaviors*. 2000; 25:769–774. [PubMed: 11023017]
- Brandon PR, Taum AKH, Young DB, Pottenger FM III, Speitel TW. The complexity of measuring the quality of program implementation with observations. *American Journal of Evaluation*. 2008; 29:235–250.
- Centers for Disease Control and Prevention. Guidelines for school health programs to prevent tobacco use and addiction. *MMWR*. 1994; 1994:43.
- Cohen J. A coefficient for agreement for nominal scales. *Education and Psychological Measurement*. 1960; 20:37–46.
- CSAP. [Retrieved May 30, 2007] Model programs: Bringing effective prevention to every community. 2001. from <http://modelprograms.samhsa.gov/model.htm>.
- Dane AV, Schneider BH. Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review*. 1998; 18:23–45. [PubMed: 9455622]
- Domitrovich CE, Greenberg MT. The study of implementation: Current findings from effective programs that prevent mental disorders in school-aged children. *Journal of Educational and Psychological Consultation*. 2000; 11:193–221.
- Drug Strategies. Making the grade: A guide to drug prevention programs (Revised Edition). Washington DC: Drug Strategies; 1999.
- Dusenbury L, Brannigan R, Falco M, Hansen WB. A review of research on fidelity of implementation: Implications for drug abuse prevention in school settings. *Health Education Research*. 2003; 18:237–256. [PubMed: 12729182]
- Dusenbury L, Brannigan R, Hansen WB, Walsh J, Falco M. Quality of implementation: Developing measures crucial to understanding the diffusion of preventive interventions. *Health Education Research*. 2005; 20:308–313. [PubMed: 15522898]
- Fleiss JL. Measuring nominal scale agreement among many raters. *Psychological Bulletin*. 1971; 76:378–382.
- Fullan M, Pomfret A. Research on curriculum and instruction implementation. *Review of Educational Research*. 1977; 52:7–30.
- Giles SM, Jackson-Newsom J, Pankratz MM, Hansen WB, Ringwalt CL, Dusenbury L. Measuring quality of delivery in a substance use prevention program. *Journal of Primary Prevention*. 2008; 29:489–501. [PubMed: 19030993]
- Gingiss PM, Robert-Graw C, Boer M. Bridge-it: A system for predicting implementation fidelity for school-based tobacco prevention programs. *Prevention Science*. 2006; 7:197–207. [PubMed: 16791522]
- Hansen WB, McNeal RB. Drug education practice: Results of an observational study. *Health Education Research*. 1999; 14:85–97. [PubMed: 10537950]
- Johnson KW. Impact of a university-government partnership in criminal justice: A linkage model for the 1990s. *The Journal of Criminal Justice Education*. 1990; 1:167–194.

- Johnson KW, Holder H, Ogilvie KA, Collins DA, Courser M, Miller B, et al. A community prevention intervention to combat inhalants and other harmful legal products among pre-adolescents. *Journal of Drug Education*. 2007; 37:227–247. [PubMed: 18047181]
- Johnson KW, Shamblen SR, Ogilvie KA, Collins DA, Saylor BL. Preventing youth's use of inhalants and other harmful legal products in frontier Alaskan communities: A randomized trial. *Prevention Science*. 2009; 10:298–312. [PubMed: 19440837]
- Lillehoj CJ, Griffin KW, Spoth R. Program provider and observer ratings of school-based preventive intervention implementation: Agreement and relation to youth outcomes. *Health Education and Behavior*. 2004; 31:242–257. [PubMed: 15090124]
- McCormick L, Steckler A, McLeroy K. Diffusion of innovations in schools: A study of adoption and implementation of school-based tobacco prevention curricula. *American Journal of Health Promotion*. 1995; 9:210–219. [PubMed: 10150723]
- McGrew JH, Bond GR, Dietzen L, Salyers M. Measuring the fidelity of implementation of a mental health program model. *Journal of Consulting and Clinical Psychology*. 1994; 62:670–678. [PubMed: 7962870]
- Melde C, Esbensen FA, Tusinski K. Addressing program fidelity using onsite observations and program provider descriptions of program delivery. *Evaluation Review*. 2006; 30:714–740. [PubMed: 17093106]
- Meyer, MA.; Booker, JM. Eliciting and analyzing expert judgment: A practical guide. Philadelphia: Society for Industrial and Applied Mathematics; 2001.
- National Center for Frontier Communities. [Retrieved May 25, 2007] List of frontier counties from 2000 U.S. Census. 2007. from <http://frontierus.org/2000census.htm>.
- NCLB. [Retrieved May 1, 2009] Safe and drug-free schools and communities act state grants guidance for state and local implementation of programs. 2004. from www.ed.gov/programs/dvpformula/guidance.doc.
- Nevo D. Expert opinion in program evaluation. *New Directions for Program Evaluation*. 1989; 42:85–93.
- NREPP. [Retrieved May 30, 2007] SAMHSA'S National registry of evidence based programs and practices: Finding interventions. 2007. from www.nrepp.samhsa.gov/find.asp.
- Ogilvie, DC.; Coulehan, HM.; Ogilvie, KA.; Johnson, KW. Think Smart: Problem solving for Alaskan youth, A curriculum to prevent substance abuse including inhalants and other harmful everyday legal products. Anchorage, AK: Akeela; 2006.
- OJJDP. Blueprints for violence prevention. 2002 from www.colorado.edu/cspv/blueprints/model/overview.html.
- Pankratz MM, Jackson-Newsom J, Giles SM, Ringwalt CL, Bliss K, Bell ML. Implementation fidelity in a teacher-led alcohol use prevention curriculum. *Journal of Drug Education*. 2006; 36:317–333. [PubMed: 17533804]
- Ridgely MS, Jerrell JM. Analysis of three interventions for substance abuse treatment of severely mentally ill people. *Community Mental Health Journal*. 1996; 32:561–572. [PubMed: 8905228]
- Ringwalt CL, Bliss K. The cultural tailoring of a substance use prevention curriculum for American Indian youth. *Journal of Drug Education*. 2006; 36:159–177. [PubMed: 17153515]
- Ringwalt CL, Ennett S, Rohrbach LA, Simons-Rudolph A, Vincus A, Thorne J. Factors associated with fidelity to substance use prevention curriculum guides on the nation's middle schools. *Health Education & Behavior*. 2003; 30:375–391. [PubMed: 19731502]
- Ringwalt CL, Pankratz MM, Jackson-Newsom J, Gottfredson NC, Hansen WB, Giles SM, et al. Three year trajectory of teachers' fidelity to a drug prevention curriculum. *Prevention Science*. 2009 (in press).
- Robertson, EB.; David, SL.; Rao, SA. Preventing drug use among children and adolescents: A research-based guide for parents, educators, and community leaders. 2nd ed.. Bethesda, MD: National Institute on Drug Abuse; 2003.
- Rohrbach LA, Dent CW, Skara S, Sun P, Sussman S. Fidelity of implementation in Project Towards No Drug Abuse (TND): A comparison of classroom teachers and program specialists. *Prevention Science*. 2007; 8:125–132. [PubMed: 17180722]

- Schinke SP, Orlandi MA, Botvin GJ, Gilchrist LD, Trimble JE, Locklear VS. Preventing substance abuse among American-Indian adolescents: A bicultural competence skills approach. *Journal of Counseling Psychology*. 1988; 35:87–90. [PubMed: 17392927]
- Schinke SP, Tepavac L, Cole KC. Preventing substance use among Native American youth: Three-year results. *Addictive Behaviors*. 2000; 25:387–397. [PubMed: 10890292]
- Shrout P, Fleiss JL. Intraclass correlation: Uses in assessing rater reliability. *Psychological Bulletin*. 1979; 86:420–428. [PubMed: 18839484]
- Sobol DF, Rohrbach LA, Dent CW, Gleason L, Brannon BR, Johnson CA, et al. The integrity of smoking prevention curriculum delivery. *Health Education Research*. 1989; 4:59–67.
- Spoth R, Redmond C, Shin C. Direct and indirect latent-variable parenting outcomes of two universal family-focused preventative interventions: Extending a public health oriented research base. *Journal of Consulting and Clinical Psychology*. 1988; 66:385–399. [PubMed: 9583342]
- Spoth R, Guyll M, Lillehoj CJ, Redmond C, Greenberg M. Prosper study of evidence-based intervention implementation quality by community-university partnerships. *Journal of Community Psychology*. 2007; 35:981–999. [PubMed: 20376336]
- Tricker R, Davis LG. Implementing drug education in schools: An analysis of the costs and teacher perceptions. *Journal of School Health*. 1988; 58:181–185. [PubMed: 3386197]
- U.S. DOE. [Retrieved May 30, 2007] Safe, disciplined, and drug-free schools expert panel. 2001. from www.ed.gov/admins/lead/safety/exemplary01/index.html.
- Yeaton WH, Sechrest L. Critical dimensions in the choice and maintenance of successful treatments: Strength, integrity and effectiveness. *Journal of Consulting and Clinical Psychology*. 1981; 49:156–167. [PubMed: 7217482]

Table 1Operational definitions of implementation quality measures for *Think Smart*

Dosage	Instructor completes or attempts to complete an activity in the session
Adherence	Categories/Description
Changes	(1) Rewords question; (2) Changes instructional detail; (3) Changes content of statement; (4) Changes logistical step; (5) Changes order; (6) Changes activity; (7) Changes question to statement
Additions	(1) New question; (2) Adds logistical step; (3) Teacher-centered example or personal anecdote; (4) Student-centered example; (5) Elaborative statement; (6) Moral lecture; (7) Adds instructional detail; (8) Adds <i>Think Smart</i> tagline
Omissions	(1) Omits instruction details; (2) Omits question; (3) Omits statement; (4) Omits logistical step; (5) Omits example; (6) Omits/truncates discussion
New delivery methods	(1) Discipline/motivational strategy; (2) Teacher led to student led; and (3) Student led to teacher led
Composite	Computed as the sum of four adherence scores
Delivery quality	Description
Praise	Instructor praises or encourages the students either enthusiastically (e.g., "Wow! Great point!") or less enthusiastically (e.g., "Good")
Accepts ideas of students	Instructor accepts a student's comment, including "parroting" of responses
Uses ideas of students	Instructor attempts to link student's comment to some aspect of the lesson (e.g. "That's like being free of bad things like drugs.")
Asks original question	Instructor asks a question not related to any previous statement or question
Asks repeated question	Instructor asks a question that they had asked previously (e.g. "Anyone else?")
Asks probing question	Instructor asks a question that refers directly to something a student has previously said

Table 2Inter-rater reliabilities for *Think Smart* dosage, adherence, and delivery skills

Indicator	% Agreement	Kappa
Dosage		
Completed/attempted	99%	.97
Adherence		
Changes	98	.90
Omissions	97	.93
Additions	96	.88
New methods	99	.95
	Pearson's r	Intra class Correlation
Delivery skills		
Enthusiastic praise	.90	.87
Less enthusiastic praise	.92	.90
Accepts ideas of students	.99	.99
Uses ideas of students	.92	.92
Asks original questions	.99	.99
Asks repeated questions	.98	.98
Asks probing questions	.89	.88

N=95

Table 3

Level of dosage and adherence to the design of the *Think Smart* curriculum

Percentage activities	Dosage	Adherence ^a				No new methods	Composite
	Completed	No changes	No omissions	No additions			
High							
80–100%	72.6%	68.1%	22.3%	50.0%	81.9%	42.6%	
60–80	15.8	24.5	36.2	33.0	10.6	46.8	
40–60	10.5	6.4	22.3	11.7	6.4	9.6	
20–40	1.1	1.1	11.7	5.3	1.1	1.1	
Low							
Total	100%	100%	100%	100%	100%	100%	
Mean %	82.2	82.0	59.3	76.0	87.6	76.2	

N=95,

^a Adherence measures are based only on activities attempted or completed

Table 4Delivery skill occurrences in the implementation of the *Think Smart* curriculum

Delivery skill	Average occurrences ^a		Average range ^b	
	Per session	(SD)	Min	Max
Praise	4	(2.5)	1	12
Accepts Ideas of Students	24	(10.4)	9	48
Uses Ideas of Students	2	(1.5)	0	5
Asks Original Question	28	(5.3)	20	42
Asks Repeated Question	7	(2.8)	3	13
Asks Probing Question	3	(2.3)	0	7

^a Mean of average delivery skill occurrences across sessions^b Minimum and maximum average occurrences across sessions

Table 5
Expert panel assessment of level and adequacy of implementation quality in comparison to other studies

	Mean rating		Mean difference ^a		<i>d</i>		Cohen <i>d</i>		% <i>p</i> < .05	
	Actual	Bootstrap	Actual	Bootstrap	Actual	Bootstrap	Actual	Bootstrap	Actual	Bootstrap
Level of implementation quality ^b										
Dosage	3.00	-	.00	.00	.00	-.01	.00	.00	-	4%
Adherence	3.00	-	.00	.00	.00	.02	.00	.01	-	5%
Delivery skills	2.47	-	-.53	-.53	-3.44 ^{***}	-3.60 ^{***}	-.89	-.93	-	98%
Adequacy of level of implementation quality ^c										
Dosage	2.63	-	.13	.12	.81	1.09	.21	.28	-	22%
Adherence	2.44	-	-.06	-.07	-.40	-.41	-.10	-.11	-	6%
Delivery skills	2.19	-	-.31	-.32	-1.67	-1.79	-.43	-.46	-	33%
Overall	2.44	-	-.06	-.07	-.40	-.40	-.10	-.10	-	6%

All *t*-tests are based on 15 degrees of freedom

^a Actual minus expected value

^b Expectations for quality were set at 3 (on a 0=much lower than average, 2=average, 3=higher than average, and 4=much higher than average scale)

^c Expectations for adequacy were set at 2.5 (on a 0=inadequate, 1=marginally adequate, 2= moderately adequate, and 3=adequate scale)

^d *p* (2-tailed)

* <.05,

** <.01,

*** <.001