Articles

Articles should deal with topics applicable to the broad field of program evaluation. Articles may focus on evaluation methods, theory, practice, or findings. In all cases, implications for practicing evaluators should be clearly identified. Examples of contributions include, but are not limited to, reviews of new developments in evaluation, description of a current evaluation study, critical reviews of some area of evaluation practice, and presentations of important new techniques. Manuscripts should follow APA format for references and style. Length per se is not a criterion in evaluating submissions.

Getting to the Bottom Line: A Method for Synthesizing Findings Within Mixed-method Program Evaluations

ANDREW McCONNEY, ANDY RUDD, AND ROBERT AYRES

ABSTRACT

Evaluators who are concerned more with pragmatics than with competing epistemologies have brought multi- and mixed-method evaluations into common practice. Program evaluators commonly use multiple methods and mixed data to capture both the breadth and depth of information pertaining to the evaluand, and to strengthen the validity of findings. However, multiple or mixed methods may yield incongruent results, and evaluators may find themselves reporting seemingly conflicting findings to program staff, policy makers, and other stakeholders. Our purpose is to offer a method for synthesizing findings within multi- or mixed-method evaluations to reach defensible evaluation (primarily summative) conclusions. The proposed method uses a set of criteria and analytic techniques to assess the worth of each data source or type and to establish what each says about program effect. Once on a common scale, simple math allows synthesis across data sources or types. The method should prove a useful tool for evaluators.

Andrew McConney • College of Education, Florida Gulf Coast University, 10501 FGCU Blvd., S., Ft. Myers, FL 33965-6565, USA; Tel: +(1) 941-590-7799; E-mail: amcconne@fgcu.edu.

American Journal of Evaluation, Vol. 23, No. 2, 2002, pp. 121–140.
ISSN: 1098-2140

INTRODUCTION

Mixed-method rather than mono-method approaches have become firmly established as common practice in program evaluation (Caracelli & Greene, 1993; Frechtling & Sharp, 1997; Greene & Caracelli, 1997; Greene, Caracelli, & Graham, 1989; Kidder & Fine, 1987; Patton, 1987; Shotland & Mark, 1987; Tashakkori & Teddlie, 1998; Weiss, 1998). Since these approaches typically include both quantitative and qualitative methods of data-gathering, the proliferation of mixed-method evaluations could be considered indicative of an ongoing trend toward pragmatism in evaluation practice (e.g., Caracelli & Greene, 1993; Greene et al., 1989; Tashakkori & Teddlie, 1998; Waysman & Savaya, 1997; Weiss, 1998). Rather than limiting method according to epistemology, evaluation pragmatists are more concerned with informing stakeholders and policy makers by using whatever type of data or method best answers evaluation questions. Pragmatists also believe that the combined use of qualitative and quantitative data may strengthen evaluations by offsetting the limitations and biases of any one method (Denzin, 1978; Greene & Caracelli, 1997; Greene et al., 1989; Jick, 1983; Rossman & Wilson, 1985; Shotland & Mark, 1987; Tashakkori & Teddlie, 1998; Waysman & Savaya, 1997; Weiss, 1998).

Accompanying the increased acceptance and use of mixed-method approaches in program evaluation, a number of evaluation writers have pointed out that using mixed methods is a challenging task, particularly with data synthesis (Jick, 1983; Kidder & Fine, 1987; Shotland & Mark, 1987; Trend, 1979; Waysman & Savaya, 1997). Jick's comment is illustrative:

> It is a delicate exercise to decide whether or not results have converged. In theory, a multiple confirmation of findings may appear routine. If there is congruence it presumably is apparent. In practice, though, there are few guidelines for systematically ordering eclectic data in order to determine congruence or validity. For example, should all components of a multi-method approach be weighted equally, that is, is all evidence equally useful? If not, then it is not clear on what basis the data should be weighted, aside from personal preference. (1983, p. 142)

Nearly two decades have passed since Jick's (1983) articulation of this question, and surely, over the years, evaluators have been repeatedly faced with the data synthesis challenge he described. Yet, in our view, the program evaluation literature has remained wanting in systematic approaches to data synthesis within individual mixed- or multi-method program evaluations for the purpose of reaching defensible summative evaluation conclusions. The lack of concrete direction we encountered led us to agree with Riggin (now Cooksy) and Caracelli that "generalities provide little guidance in the current call for synthesis which focuses on how evidence from multiple perspectives is put together for the purpose of deriving a coherent assessment" (Riggins & Caracelli, 1994, p. 139).

Thus, the purpose of this paper is to begin to respond to the lack of guidance we have found by offering evaluators one method for synthesizing the findings from program evaluations that utilize mixed- or multi-method approaches. That is, we here propose a method of "results synthesis" within mixed- or multi-method program evaluations.

Although we hesitate to demarcate the method offered by saying what it is not, at the same time we believe that doing so may mitigate potential misunderstanding for readers. In this spirit, the synthesis-of-findings method proposed is not meta-analytic in the classic sense of the term (e.g., Glass, 1976; Kulik & Kulik, 1989). That is, the synthesis method offered

does not suggest a way of looking *across* multiple evaluations for the purpose of synthesizing findings related to an issue of research or evaluation interest (e.g., class size reduction, use of concept mapping as an instructional strategy). Rather, it proposes a method of looking *within* an individual program evaluation to synthesize the findings that derive from multiple or mixed methods of data-gathering.

Neither is the synthesis method proposed meta-evaluative in the well-accepted usage of that term (e.g., Scriven, 1969; Stufflebeam, 1974). That is, we have not devised a method for assessing and judging the overall merit, worth, or value of evaluations. Rather, we propose a procedure by which the quality of each line of evidence gathered within individual mixed- or multi-method program evaluations may be assessed. Quality of evidence assessments, in combination with judgment-based estimates of what each line of evidence says about program effect, can then be aggregated across diverse data-gathering methods to arrive at reasoned estimates of program effectiveness. We suggest that such estimates of program effectiveness have the potential to serve evaluators as important checks on the summative conclusions they may be required to reach. Further, given our experience, we suggest that the synthesis of mixed-method evaluation findings proposed here will be of particular service when data divergence is evident and unlikely to be resolved (e.g., some lines of evidence suggest that the evaluand is effective, while others suggest that it is not).

Briefly then, the aim of this paper is to first describe and then demonstrate a flexible, judgment-based technique by which evaluators using mixed-method approaches might synthesize their findings to reach evaluative conclusions. The data synthesis method (1) rates (and thereby weights) each piece of qualitative or quantitative evidence using "criteria of worth"(CoW) aligned with the aims of the program under evaluation; and, once evidences are on a common scale, (2) provides a method whereby they can be aggregated to assist evaluators in reaching a defensible conclusion. The method is suggested as appropriate for mixed-method summative evaluations that call for conclusions that routinely (1) result in a "thumbs-up" or "thumbs-down" for the evaluand; or (2) potentially result in a change in policy underlying the evaluand. Further, the method is suggested as useful in mixed-method situations where the findings of different data-gathering methods appear in conflict, that is, when the data synthesis challenge portrayed in the shaded area of Figure 1 occurs.

Result Possibilities	Line of evidence #1 (e.g., quantitative)	Line of evidence #2 (e.g., qualitative)	Conclusion (program merit or worth)
Example 1	+	+	+
Example 2	—	—	—
Example 3	+	—	?
Example 4	—	+	?

Figure 1. Conclusion Possibilities after Synthesis-of-findings from Mixed- or Multi-method Evaluations.

METHOD

Having provided the terms of reference that delineate the method we propose, we now turn to the method itself. This part of the paper is divided into two sectors: (1) a step-by-step generalized description of the method in the abstract; and (2) a fairly detailed example of how the method was applied to the findings of a "real-life" mixed-method program evaluation.

In the Abstract

This synthesis-of-findings method is intended to assist program evaluators in reaching reasoned conclusions within individual mixed- (quantitative and qualitative lines of data) or multi-method (multiple lines of data which may be all quantitative or all qualitative) evaluations. We suggest that this method should be particularly helpful in summative situations in which some methods or data provide positive evidence about the program, and some provide negative evidence about the program; that is, data divergence is apparent.

The proposed data synthesis method consists of four steps that are given first in brief, and then in more detail:

1. Rate the program's effect according to each evidence set. That is, determine and record ratings of the direction and magnitude of the program's effect according to what each source, type, or set of evidence is saying about the program under evaluation—positive effect, no effect, or negative effect.
2. Rate each evidence set's worth. That is, determine and record program-specific quality ratings for each source, type, or set of evidence according to "CoW" aligned with the particular program.
3. Combine program effect ratings with "CoW" ratings, according to the program goal or outcome being examined.
4. Aggregate combined ratings to arrive at a summary program "effectiveness estimate."

Step 1. After analyzing each data set in relation to its associated program goal or outcome, interpret and record the direction and magnitude of the program's effect. That is, what do the data gathered over the course of the evaluation have to say about the effect of the program, on a method-by-method basis? On a method-by-method basis, determine if the message is unequivocally positive? Unequivocally negative? Slightly positive? Slightly negative? Neutral?

This first step will probably need to be done on an outcome-by-outcome basis, or a goal-by-goal basis, as it is more often than not the case that programs have multiple goals and objectives, or multiple targeted outcomes. Also, it is probable that the data collection strategies for the evaluation have been designed around individual program goals, objectives, or targeted outcomes. We suggest the rubric given in Figure 2 for recording effect ratings for the program.

Program effect ratings may be achieved in a number of ways. For example, if an individual is conducting the evaluation, then it is probably wise to have a colleague review the raw data and analytic strategies used, as well as the program effect ratings assigned by goal/outcome and data source. If, on the other hand, a team of evaluators conducts the evaluation, then it is more likely that the team members will take advantage of the opportunity to cross-check data, analyses, and findings for each goal or targeted outcome of the program. Still, whether it is an individual evaluator or an evaluation team, it would be advisable at this stage to involve a stakeholder—or external advisory panel to collaboratively review the data gathered around

Data Set/Source	Program Effect (direction and magnitude)				
	Large, negative effect (-2)	Small, negative effect (-1)	No discernible effect (0)	Small positive effect (+1)	Large positive effect (+2)
Goal/outcome #1					
Data source/set #1					
Data source/set #2					
Goal/outcome #2					
Data source/set #1					
Data source/set #2					
Data source/set #3					

Figure 2. A Rubric for Recording Ratings of Program Effect, by Goal and Data Source, in Mixed- or Multi-method Evaluations.

each program goal/outcome, the analyses conducted, as well as what each data set is saying about the program, i.e., the program's effect rating based on that data.

Additionally, if the data collection effort has included gathering quantitative data that allow the calculation of effect size estimates (i.e., there are both treatment [program] and control or comparison group results available), then by all means use this procedure as an aid in completing Step 1. Cohen (1977), for example, has provided benchmarks in standard deviation units that could be used to categorize calculated effect size estimates as small, medium, or large (small = 0.25, medium = 0.50, and large = 1.00). If the evaluator were to use these benchmarks, we note that the effect size categories given in Figure 2 would need to be expanded to reflect Cohen's scheme. In addition, however, we would encourage program evaluators to focus on the practical significance of any computed effect size, rather than rely solely on suggested benchmarks in the abstract. For example, if an elementary school reading program resulted in an effect size of 0.50 on some criterion of interest (indicating that the typical student in the "treatment" group placed at the 69th percentile, as compared to the typical comparison group student at the 50th percentile) we might reasonably judge this to be a large effect given the complexity of successfully introducing programs in the schools. That is, evaluators' and stakeholders' contextually grounded judgments about the magnitude of program effect should not be overridden by benchmarks offered as guidelines about effect sizes in the abstract. To summarize, then, the purpose of Step 1 is to examine each set or source of evidence to determine what each says about the direction and strength of the program's effect, in relation to the particular program goal or outcome being studied.

Step 2. Rate each data set, or source, or type of evidence against agreed-upon "CoW" for the program under evaluation. Criteria could be arrived at collaboratively with program stakeholders, especially those contracting the evaluation, as well as those potentially most affected by the program, if possible. At a minimum, the CoW used for the synthesis exercise should be public, and described in sufficient detail so that they can be examined and understood by others with reasonable effort. In effect, these "CoW" ratings will form the weights for each piece of evidence brought to bear in judging the program's effectiveness (most likely, on a goal-by-goal basis).

Below we provide seven CoW that we suggest evaluators may use to screen sources or sets of data. This list of criteria is provided simply as an example of what might be considered

when thinking about assessing the quality of sets or sources of evidence. It is probable that no one list of CoW will be sufficiently exhaustive that it captures every conceivable evaluation or program scenario, and thus it is likely that each evaluation will use different criteria. Yet, we are not uncomfortable with this likely lack of standardization; rather, we believe that two related processes are of critical importance here. The first is recognizing that not all data are created equal (data are likely to vary in quality depending on source, method, instrument, administration, etc.); the second involves collaboratively gaining consensus on what criteria are appropriate and useful for assessing the quality of data within a particular mixed- or multi-method evaluation. We believe that these processes hold strong potential for enhancing ownership of the evaluation and thus use of the evaluation findings as well.

In the following sample list, the CoW are thought to be at a level general, yet specific, enough to be useful across a range of program evaluation possibilities. Also, the criteria offered here are compatible, though they do not map one-to-one with Stufflebeam's criteria of technical adequacy and utility described in his seminal paper on meta-evaluation (Stufflebeam, 1974). However, the criteria we offer focus solely on issues related to the quality of data gathered, whereas those provided by Stufflebeam, while including data quality, also address issues of evaluation conduct, utility, and cost effectiveness. In addition, the CoW list in form and function may even more closely resemble a criteria of merit list (comlist) as conceptualized by Scriven (2000). In the current case, however, we prefer the term worth, which carries the notion of value in context; that is, the value or excellence of any datum must be assessed against criteria important in the context of a particular evaluation. Readers will further note that where applicable in the criteria we offer, we have attempted to include data quality considerations or concepts from both positivist/postpositivist and interpretative/constructivist epistemological traditions (e.g., Mertens, 1998).

Accurate/credible?[1] What evidence exists that the data are accurate? That is, are the data likely to be valid? What evidence is there that the data derive from an instrument that measures the construct it claims to measure? Alternatively, to what degree are the data credible? That is, given the length of contact, persistence, and intensity of the data-gathering effort as well as the vantage point and credentials of the data source, to what degree are these data believable (plausible, convincing) for (a) intended users (those implementing) of the program, (b) relevant policy makers who may determine the continued and/or expanded use of the program, and, importantly, (c) intended beneficiaries of the program, i.e., those stakeholders on whom the program is most squarely targeted?

Reliable/dependable? What evidence exists that the data are reliable? That is, are the data trustworthy? Does the data-gathering instrument provide stable, consistent results across raters, subjects, and/or time? Alternatively, to what degree are the data provided dependable? Are they provided by a contextually knowledgeable source, that is, some person or persons who have the experience and expertise to make the required assessment or interpretation? As well, is there a transparent, auditable trail of how interpretations and changes in interpretations were developed over time?

Close to those impacted? From whom are these data collected? That is, are these data collected directly from those primarily impacted by the program (e.g., students)? Are these data collected from a source "once-removed" from those potentially impacted (e.g., parents or teachers)? Or are the data collected from sources more than "once-removed" from those

potentially impacted (e.g., a school administrator offering assessment of students' improvement in reading based on her observations or conversations with teachers)?

Relevant? This criterion bridges Stufflebeam's (1974) "relevance" and "importance" criteria by asking for the degree to which these data are well-aligned with and important for answering the questions asked of the evaluation, the goals and objectives of the program, and/or the claims of the program purveyors. For example, if the program purports to improve academic achievement, does this source provide data about change in academic achievement? As noted by Stufflebeam (1974, p. 7), "Relevance is determined by comparing each datum to be gathered with the questions to be answered."

Representative? This criterion is very similar to Stufflebeam's (1974) "scope," and asks the degree to which the scale, or scope, of assessment represented by these data match the scale of the intended use of the program. In other words, to what degree is the scale of this data source or type aligned with the scale, or potential scale, of the program? For instance, if a "reading improvement" program is intended potentially for all elementary-age school children in a state, how well do the data represent, in quantity and in character, elementary-age school children in the state?

Understandable? To what degree are these data comprehensible (explicable for, reasonable) for (a) intended users (those implementing) of the program, (b) relevant policy makers who may determine the continued and/or expanded use of the program, and, importantly, (c) intended beneficiaries of the program, that is, those stakeholders on whom the program is most squarely targeted?

Well-collected? To what extent are these data free from bias potentially associated with instrument administration and/or the collection method used? For example, if the data derive from paper-and-pencil instruments, was the instrument administered appropriately? Or, if the data were collected from a focus group or individual interview, was the interview conducted according to accepted guidelines? For example, were interviewers trained and experienced? Were those in potential positions of power relative to the interviewees excluded from the interview (e.g., school principals excluded from a focus group composed of teachers)? Overall, were these data collected skillfully and with attention to minimizing possible threats to their integrity?

To reiterate, the purpose of Step 2 is to assess the quality of each set or source of data using collaboratively agreed CoW. We expect that some will disagree with the composition of the above list, and that some may dispute the criteria meanings we have offered. This is welcomed because such a challenge can only broaden and improve the criteria we use to judge the quality of evidence gathered within evaluations. More importantly, the main message here is not the specific content of the CoW list offered, but our advocacy for the process of systematically and collaboratively assessing each data set so that the worth of each can be compared relative to others.

Step 3. Using the ratings generated in Steps 1 and 2, write the aggregation (synthesis) equation for each program goal or outcome under consideration. The equation will be of the form given in Figure 3. In essence, this equation says that an estimate of program effectiveness, for any program goal or outcome, is obtained by summing the products of program effect ratings and data worth ratings, applicable to a particular outcome or goal.

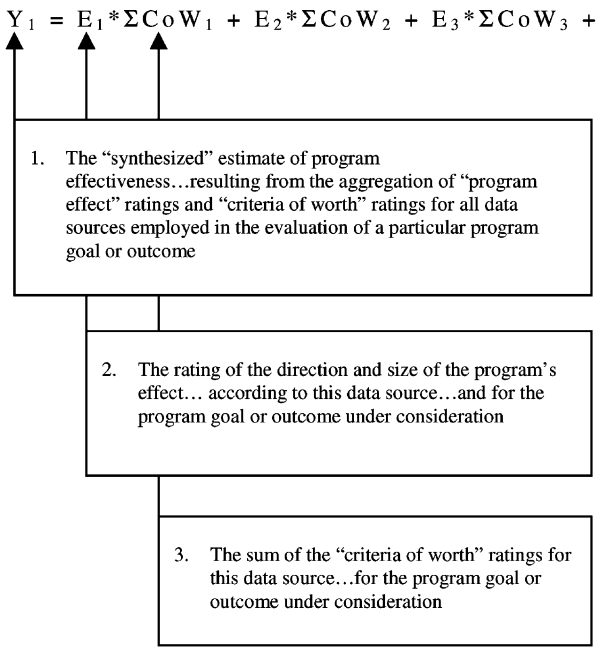

Figure 3. An Equation for Synthesizing Program Effect and CoW Ratings in Mixed- or Multi-method Evaluations.

This procedure will result in scoring the program (by outcome) on a closed program effectiveness scale, the end-points of which are determined by the number of CoW used in screening the data sets gathered, as well as the number of data sources applied in evaluating a program. That is, the procedure will result in a program effectiveness estimate for each program goal or outcome evaluated.

For the following example, we have chosen a program evaluation in which we gathered five distinct data sets, resulting from five different data collection methods. As a result of the Step 1 exercise, each data set carries an effect rating ranging from -2 (large negative effect) to $+2$ (large positive effect). Similarly, we employed five CoW, each of which can range from $+1$ (low) to $+3$ (high). This means that the total CoW rating for each set of evidence could range from a low of 5 (i.e., 5×1; low worth data) to a high of 15 (i.e., 5×3; high worth data). Thus, the product of each "effect rating" and "CoW rating" could range from a low of -30 (negative effects, worthy data) to a high of $+30$ (positive effects, worthy data) for each data set or source. Summed over five data sets, this would mean a program effectiveness estimate that could range from a low of -150 (negative effects, worthy data; that is, a negatively effective program) to $+150$ (positive effects, worthy data; that is, a positively effective program).

Step 4. After Steps 1–3 have been completed for each program goal or targeted outcome of the program under evaluation, goal-wise effectiveness estimates are averaged to produce one overall program-wise effectiveness index. This in no way implies that the program learning associated with the rich mixed data gathered around each program goal should be ignored or discarded. Neither does it mean that the evaluator should not further explore or attempt to

reconcile divergence in the evaluation's findings. We advocate for both of these, but at the same time offer this check on program effectiveness as one "bottom line" measure if the evaluator is charged with providing a summative program-wise conclusion.

Also at this stage, some measure of the program's fidelity of implementation should be factored into the synthesized estimate of program effectiveness. This is conceptually similar to the notion of correcting for attenuation whereby observed correlation coefficients may be "corrected" based on measures of unwanted or artificial variation such as the reliability of an instrument (Hunter & Schmidt, 1994). It is also conceptually similar to the idea of correcting for bias in estimating effect sizes (Hedges, Shymansky, & Woodworth, 1989). Thus, if program evaluators, developers, and/or users estimate that the evaluand is being implemented with 75% fidelity to the ideal, then the program effectiveness estimate resulting from the previous steps could be divided by 0.75 (i.e., mediated by the less than ideal implementation). By dividing we are essentially saying that if the program were to be faithfully implemented, it likely would be rated that much more effective. If, however, it turns out that fidelity of implementation varies according to program goal or targeted outcome, then the fidelity measure should be factored into the picture at the level of the program goal being considered.

Real-life Example

The example to be used here is the questions-based, mixed-method program evaluation of the Save Our Schools (SOS) pilot program conducted by ERGO for the state of Euphoria, under contract to the Office of Special Education, Department of Education. (Although the evaluation is real, we have chosen to use fictitious names to (1) maintain the focus of this paper on the synthesis method proposed, and (2) protect program and program evaluation stakeholders from any potential but unintended harm.)

Purpose of the evaluation study. Simply stated, the primary purpose of this third-party evaluation was to determine the effectiveness of the SOS Model School program as implemented in a pilot program in three elementary schools beginning in February 1998, extending to 16 additional elementary schools (a total of 19 pilot schools) for the 1998–1999 school year, and continuing in 17 pilot elementary schools for the 1999–2000 and 2000–2001 school years. For this evaluation, program "effectiveness" was operationally defined in alignment with the original state Senate Bill that authorized the program and the claims made for it by its developers and marketers, and in conjunction with state Department of Education staff, as:

1. Improvement in student achievement in reading, writing, and mathematics.
2. Decreases in referrals for special education assessment and/or services.
3. Improvement in students' behavior.
4. Improvement in students' rates of English language acquisition (for students whose primary language is other than English).
5. Improvement in school attendance rates.

Save Our Schools program. According to its literature, the SOS program is an education program that uses a combination of structured curricula in the form of classroom modules, and in school Learning Center to teach and develop important learning abilities for students. The SOS program focuses on 26 intellectual abilities that are claimed to be most critical to effective learning (e.g., the abilities needed to acquire, store, evaluate, and use information). These 26

abilities are taught in activities grouped around learning preparation, learning enhancement, and learning remediation. Learning preparation is addressed in classroom exercises that are designed to take place for 15–20 minutes per school day. Similarly, learning enhancement is also accomplished through classroom activities. In both cases, SOS classroom modules are articulated in difficulty through 8–12 exercises, and all materials are provided to the classroom teacher with no teacher preparation required. Learning remediation, on the other hand, is addressed in the Learning Center where students are assessed in terms of cognitive abilities, perceptual skills, and sensory-motor skill integration. Students' learning ability deficiencies are diagnosed and treatment plans are provided either on a group basis (grades K-2) or on an individual basis (grades 3–5/6). Students participate in Learning Center activities ideally for 30 minutes, twice each week. The SOS program (classroom modules and Learning Center activities) are designed as a "treatment" to be completed by students within 7 months, that is, the time span of a normal school year.

According to the SOS model, classroom exercises and an SOS Learning Center housed in a participating school lead to improvements in the achievement levels of all students. This improvement extends as well to students with learning disabilities and other special needs as well as to students whose behavior is problematic in school settings. Thus, the developers and providers of the SOS program state that schools can expect the following outcomes:

- Increased academic performance.
- Decreased special education referrals.
- Decreased disciplinary referrals.
- Increased rates of English language acquisition.
- Increased school attendance.

Questions addressed by the evaluation. ERGO's third-party evaluation of the SOS program was designed to study its effectiveness for students in participating elementary schools, with regard to academic performance, special education referrals, behavior referrals, school attendance, and English language acquisition for students who have a first language other than English.

To assess SOS program effectiveness, the evaluation team addressed five key questions:

1. Are there differences in students' academic performance in reading, writing, and mathematics between schools participating and not participating in the SOS program?
2. Are there differences in levels of special education referral between schools participating and not participating in the SOS program?
3. Are there differences in levels of behavior referral between schools participating and not participating in the SOS program?
4. Are there differences in language acquisition rates for students with ESL between schools participating and not participating in the SOS program?
5. Are there differences in student attendance rates between schools participating and not participating in the SOS program?

Procedures used to gather, analyze, and interpret data. To answer the five questions posed, the evaluation team used a core quasi-experimental design supported by selected case studies, teacher surveys, focus group interviews, and on-site observations. The general evaluation design is depicted in the schematic in Figure 4. Thus, a variety of quantitative and qualitative data were gathered using a variety of collection methods. First, and central to the evaluation,

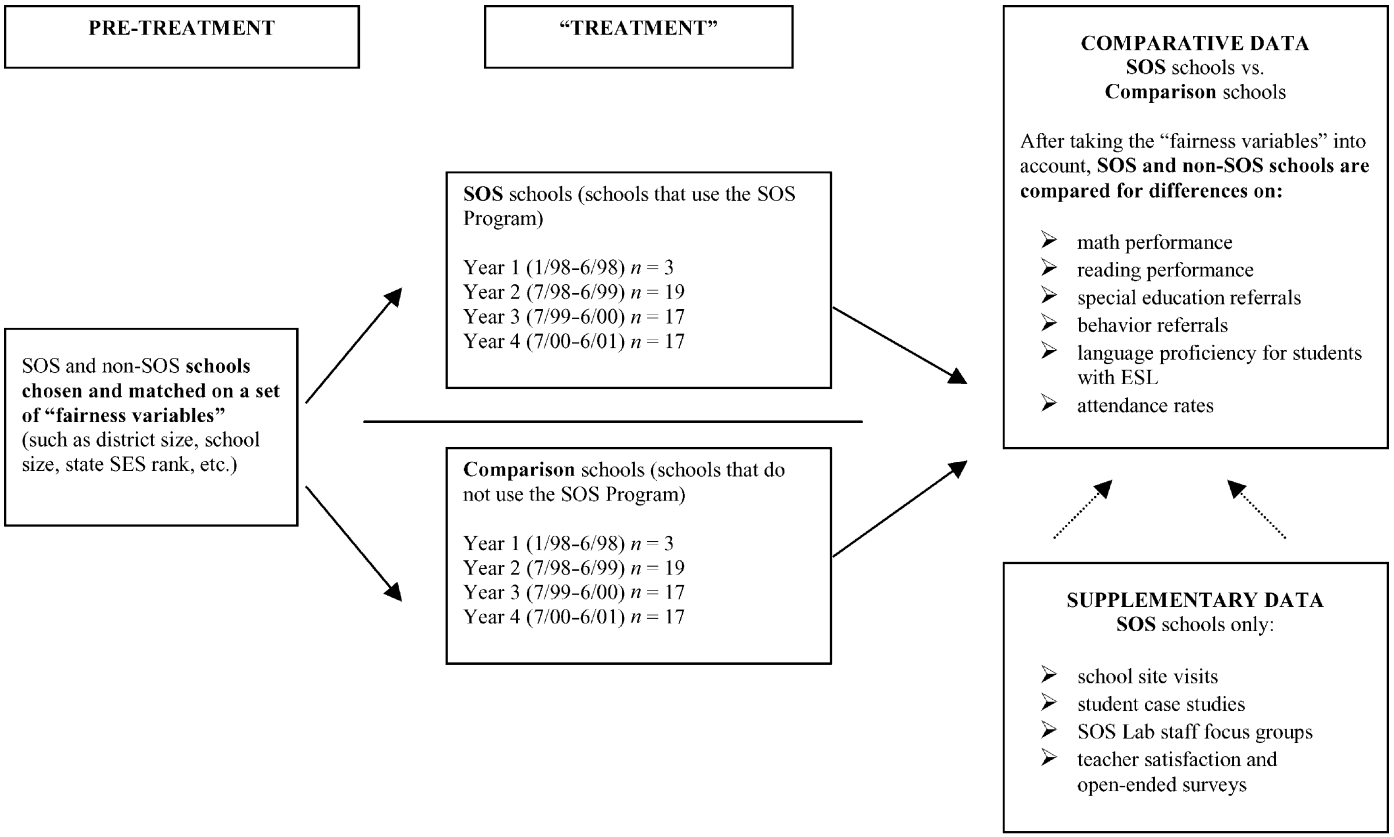

Figure 4. Overall Quasi-experimental Design for Evaluation of the SOS Program.

student achievement data in reading, writing, and mathematics from state assessments at benchmark grades 3 and 5 were collected with the cooperation of the Department of Education. Second, data collection instruments were developed to collect school-based quantitative data on student referrals for special education eligibility assessment, student referrals for inappropriate behavior, numbers of students entering or leaving English as a second language (ESL) services, school attendance rates, and levels of teacher satisfaction with the SOS curriculum.

Both state assessment data and the data provided directly by participating schools were used in graphical and statistical analyses to provide answers to the five key evaluation questions posed previously. To "level the playing field" as much as possible before comparisons were made, comparison schools were carefully selected to match the SOS pilot schools using variables such as school socioeconomic status (SES) rank, school size, and previous performance on state assessments (by grade and subject).

Four qualitative methods were used to supplement and support the quantitative lines of data. The first qualitative method employed was the "school site visit." Participating SOS school visits included an initial site visit to each school by the evaluation team, during which the team interviewed the school's principal and the SOS Learning Center staff, and inspected the school facilities designated for use as the SOS Learning Center. Over the course of each school year, two members of the evaluation team visited each SOS pilot school three times (fall, winter, and spring).

The second type of qualitative method used was the "case study." By providing detailed descriptions of the school and home backgrounds of selected children from multiple perspectives, case studies allowed the evaluation team to study how the program (particularly the Learning Center) worked for individual students over time, with particular focus on changes in the outcomes targeted by the SOS program. In all, 22 case studies were conducted with students from schools using the SOS program with sources of data including teachers, special education and other resource specialists, parents, and the students themselves.

The third qualitative method used in the evaluation was the "focus group interview." Focus groups provide good opportunities to learn directly from a group of stakeholders, in this case primarily SOS school staff, on questions of interest to the program evaluation. Typically, the evaluation team posed questions pertaining to: (1) program training and follow-up support; (2) administrative support and program fit; (3) parent and community reactions to the pilot program; (4) classroom teacher implementation of and reaction to the SOS modules; (5) SOS school staff reaction to this third-party evaluation; and, most significantly; and (6) perceived SOS program effects and/or impacts for students.

The fourth qualitative method used in this program evaluation was an "open-ended teacher survey." The evaluation team designed a survey for teachers largely comprised of open-ended questions aligned with the dimensions of student outcomes addressed by the evaluation.

The data synthesis challenge. In each of the 3 years we provided annual reports, as well as in the conclusion of the 4-year pilot program (our summative evaluation report to the Department of Education), we concluded that the SOS program was less than effective. We reported that the program had not achieved its claimed school-wide improvements, the conceptual and contractual bases on which the program had been adopted, for the pilot schools involved. These conclusions were based on the extensive data gathered and analyzed, which in turn were aligned with the school-wide improvement claims made for the SOS program.

However, each year we experienced a certain sense of disquiet because each year data divergence was apparent. That is, the messages provided by the data (findings) were consistently

positive or negative regarding program effectiveness depending on the type or source of data examined. The large-scale standardized state assessment data, and the school-wide quantitative data both provided consistently neutral or negative findings on program effectiveness. On the other hand, the site interview, focus group, and case study (primarily qualitative) data provided consistently neutral-to-positive messages about program effectiveness. Thus, each year we found ourselves prodded to ask the question: "Are we being fair, in our data synthesis and reporting, to the program developers, program providers, and school staffs who see the program as worthwhile and effective?" That is, were we falling prey to the trap noted by Jick (1983), and simply filtering the findings through lenses based on personal preference, or training, to arrive at conclusions in line with the data source that *we* valued most highly? It was this dilemma, continuing for more than 3 years, that led us to seek out and subsequently develop a method of defensibly synthesizing findings from mixed-method evaluations.

Program Goal: Academic Achievement

Step 1	Program Effect (direction and magnitude)				
Data Set/Source	Large, negative effect (-2)	Small, negative effect (-1)	No discernible effect (0)	Small positive effect (+1)	Large positive effect (+2)
1. Statewide assessments			√		
2. Teacher surveys				√	
3. Focus groups					√
4. Site visit interviews					√
5. Case studies				√	

Step 2	Criteria of Worth (low=1, medium=2, high=3)					
Data Set/Source	Accurate/ Credible?	Close to those Impacted?	Relevant?	Representative?	Well-Collected?	CoW total
1. Statewide assessments	**3**	**3**	**2**	**3**	**2**	**13**
2. Teacher surveys	**3**	**3**	**3**	**2**	**2**	**13**
3. Focus groups	**1**	**2**	**2**	**1**	**3**	**9**
4. Site visit interviews	**1**	**2**	**2**	**1**	**2**	**8**
5. Case studies	**3**	**3**	**2**	**1**	**2**	**11**

Step 3

$Y_{\text{academic achievement}} = 0*13 + 1*13 + 2*9 + 2*8 + 1*11 = 58$

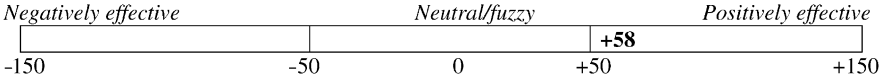

Figure 5. Three-step Synthesis Method Applied to Academic Achievement for the SOS Program.

Program Goal: School Behavior

Step 1	Program Effect (direction and magnitude)				
Data Set/Source	Large, negative effect (-2)	Small, negative effect (-1)	No discernible effect (0)	Small positive effect (+1)	Large positive effect (+2)
1. School data			√		
2. Teacher surveys			√		
3. Focus groups				√	
4. Site visit interviews				√	
5. Case studies				√	

Step 2	Criteria of Worth (low = 1, medium = 2, high = 3)					
Data Set/Source	Accurate/ Credible?	Close to those Impacted?	Relevant?	Representative?	Well-Collected?	CoW total
1. School data	**2**	**3**	**3**	**3**	**2**	**13**
2. Teacher surveys	**3**	**3**	**3**	**3**	**2**	**14**
3. Focus groups	**2**	**2**	**3**	**1**	**3**	**11**
4. Site visit interviews	**1**	**2**	**1**	**1**	**2**	**7**
5. Case studies	**3**	**3**	**2**	**1**	**2**	**11**

Step 3

$Y_{behavior} = 0*13 + 0*14 + 1*11 + 1*7 + 1*11 = 29$

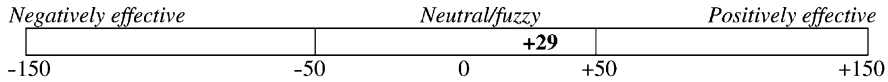

Figure 6. Synthesis Method Applied to School Behavior for the SOS Program.

Applying the data synthesis method. To illustrate use of our synthesis method in a "real-life" program evaluation, we have chosen two program goals for the SOS program under evaluation: (1) academic achievement, and (2) school behavior. Similarly, for the sake of parsimony, we have chosen five of the seven CoW presented previously that seem most appropriate to this particular program evaluation: (1) accuracy/credibility, (2) closeness to those impacted, (3) relevance, (4) representativeness, and (5) well-collected. In Figures 5-7 we present our ratings of program effect (Step 1), and on CoW (Step 2), for each data set or source, by program goal. Following the ratings for each program goal, we present the applicable synthesis equation (Step 3).

In synthesis Steps 1 and 2, each data set or source is rated in terms of what each says about program effect, and in terms of the quality of the evidence. Once each set of data applicable to each program goal has been rated, program effectiveness with respect to each targeted outcome

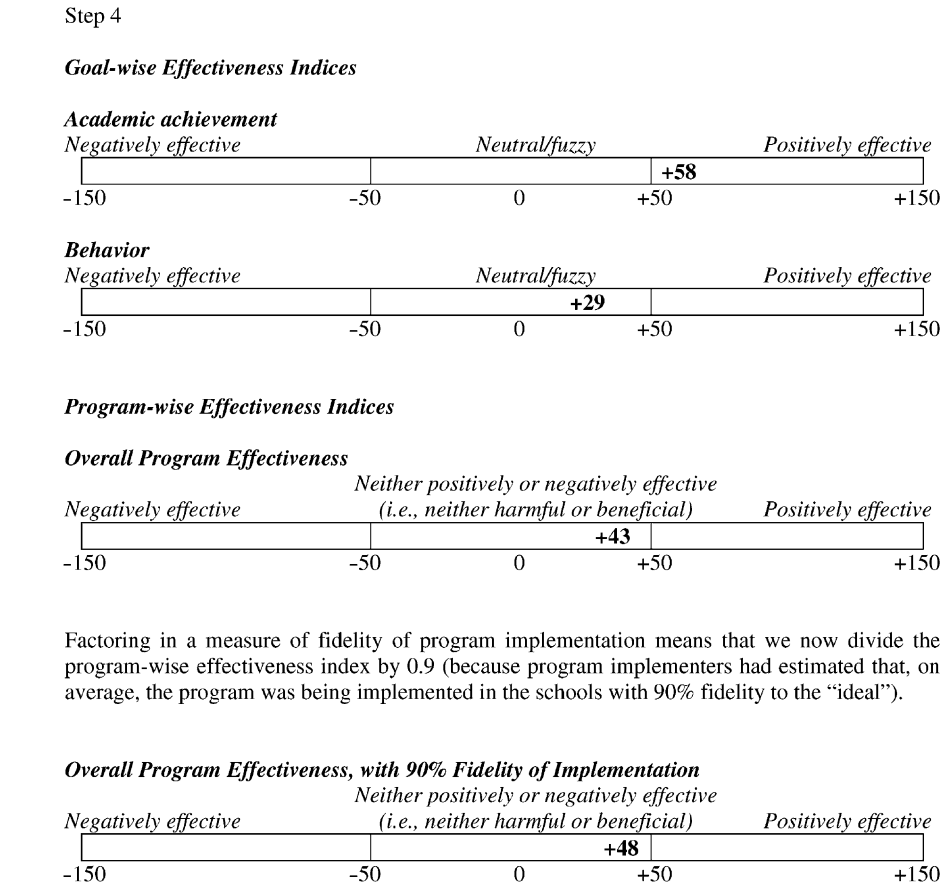

Figure 7. Program-wise Effectiveness as an Average of Goal-wise Effectiveness Indices.

(goal-wise effectiveness) can be displayed individually and/or graphically as shown in Figures 5 and 6. Alternatively, if the evaluator(s) finds a useful purpose served by aggregating effectiveness indices across program goals a defensible index of program effectiveness (program-wise effectiveness) may be calculated by averaging the goal-wise effectiveness indices, as shown in Figure 7.

Finally, as also shown in Figure 7, in estimating program-wise effectiveness the evaluator(s) should consider taking into account some measure of the fidelity of the program's implementation. For the current case, we believe that the SOS staffs in the schools are in the best position to judge the degree to which the program has been faithfully implemented. Thus, in each year of the evaluation, toward the end of the school year, we asked program staffs to look over the year and estimate on a scale ranging from 1 (no fidelity to the ideal) to 10 (perfect fidelity to the ideal) the implementation of SOS at their schools. In general over 4 years, fidelity of implementation estimates varied in a narrow range from 7 to 9. Purely for illustration purposes, then, in Figure 7 we chose a fidelity of implementation estimate of 9; that is, in the views of those most centrally connected to the SOS program as it operated over 4 years in this state's schools, the program was implemented at about 90% fidelity in relation to the ideal.

As seen in Figure 7, the data synthesis method results in a program-wise effectiveness index of +48. That is, by this method, the SOS program is rated "neither positively nor negatively effective."

DISCUSSION

Thirty years ago, as the evaluation transdiscipline was gaining traction, one of the field's definers, writing on program evaluation in education, noted:

> They [evaluators] sometimes sound as if they stopped in midstride. Why don't they sum the evaluation up and just say the project was, for example, a practically unqualified failure, as it often is? Because of the possibility of hurting feelings? As an evaluator, that's irresponsible; your obligation is to the funding agency, and ultimately the taxpayer, and not to the educational project being evaluated. Medical and industrial researchers have their feelings hurt all the time—most innovations are unsuccessful—but we don't want dangerous or useless drugs or devices on the market, hence the regulatory commissions. The educational evaluator needs to remember he's the public protector in this area. It's his task to condense the mass of data into one word: good, or bad. Sometimes this really is impossible, but all too often the failure to do so is simply a cop-out disguised as or rationalized as objectivity, or description rather than prescription! (Scriven, 1971, p. 53)

We could not agree more strongly nor say this any more eloquently than has Scriven. At the same time, we are committed to Program Evaluation Standard A10, Justified Conclusions, which reads as follows: "The conclusions reached in an evaluation should be explicitly justified, so that the stakeholders can assess them" (Joint Committee, 1994, p. 177). We are further compelled by the analysis of the illustrative case given for A10, where it is noted: "they [the evaluators] should have formulated their conclusions and recommendations to reflect *all the findings* [our emphasis] of the field test—not just those reflecting student performance" (Joint Committee, 1994, p. 179).

However, while we are informed by standard A10 and other Accuracy and Utility standards in thinking about differentiating the quality of data, we did not find the concrete procedural or conceptual guidance we sought for systematically synthesizing the findings that result from mixed- or multi-method evaluations. Thus, with the aim of adding to the evaluator's repertoire of tools that provide guidance for tackling the synthesis task, but by no means as a replacement for good judgment, we have proposed a data synthesis method by which evaluators might systematically, reasonably, usefully, and defensibly merge the findings resulting from different data-gathering methods or data sources within mixed- or multi-method program evaluations. We suggest that this method now delivers one means for arriving at a coherent assessment, and will provide an aid to evaluators in meeting standard A10, especially in situations that (1) require the evaluator(s) to reach "bottom line" (summative) conclusions about program effectiveness, and (2) confront the evaluator(s) with divergent findings depending on the data method or source.

We had been prompted to think about, research, and ultimately develop such a method by real need—a 4-year program evaluation that had repeatedly presented both of the situations noted above; that is, a summative judgment is required, and findings diverge depending on data source. That is, the qualitative data, (e.g., focus groups with program staffs and teachers, and case studies on selected students) had suggested that SOS is an effective program for improving various outcomes for elementary school children, including their academic achievement.

Conversely, quantitative data had clearly shown that students had not benefited to any significant degree on these same outcomes as a result of the SOS program. Our purpose in applying the data synthesis method in this example was not to reconcile the divergence between qualitative and quantitative data, but rather to let that divergence stand while at the same time informing and reaching a defensible summative judgment. As noted earlier, this does not mean that data divergence should not be pursued with a view to reconciliation, and we strongly advocate that evaluators and stakeholders meet to pursue whatever divergence is apparent (especially what might be learned from it).

As a result of applying the synthesis method to the evaluation example given, we arrived at a program-wise effectiveness index of +45 on a scale ranging from −150 to +150. On the program effectiveness continuum applicable to this evaluation, we interpret this score to mean that the program is "neither positively nor negatively effective," and we are encouraged that this result is consistent with what we have reported. Further, however, we hasten to note that we view this method and the effectiveness index produced by it not as the "final word" on program effectiveness, absent context or judgment. Rather, we view the effectiveness index produced as a guide or aide to evaluators that should be used very much in conjunction with their knowledge of and judgment about the evaluand. And we reemphasize that in the final analysis, it is not the particulars of the example detailed here that are important, but rather the systematic process by which each evidence source or type of data is rated for effect and quality, placed on a common scale, and then aggregated to provide the basis for a defensible conclusion.

Limitations

Despite our advocacy for the synthesis method described in this paper, and our illustration of its application to the SOS program, we emphasize its limitations.

First, the method applies only to a limited subset of program evaluation possibilities. As previously mentioned, the synthesis method we have developed is not appropriate for all types of program evaluations. Not all evaluations require summative conclusions, and not all mixed-method evaluations produce divergent findings. Also, the *a priori* purpose of a mixed-method design, that is, initiation, expansion, development, complementarity, or triangulation, may influence how the evaluator chooses to analyze and subsequently synthesize data. Caracelli and Greene have argued that an integrative approach to data analysis is well suited for the purposes of initiation, expansion, development, and complementarity, whereas the purpose of triangulation "requires independence of methods through data analysis and interpretation . . . [because] arguments for convergent validity of findings from different methods are stronger when such independence can be claimed" (Caracelli & Greene, 1993, p. 204). While we agree with this statement, we do not share the view that evaluation purpose necessarily delimits the analytic approach. Initially, we began with a parallel approach to data analysis to serve both triangulation and expansion purposes, but once data divergence became apparent the integrative synthesis-of-findings method offered here was developed and applied as a check on the summative conclusions reached through the examination of essentially parallel data (for another example of how an evaluation used a parallel approach for purposes other than triangulation, see Waysman & Savaya, 1997). Still, the caveat remains that this method of data synthesis serves only a subset of evaluation scenarios.

Second, the fidelity of implementation factor should be applied according to the evaluator's knowledge of the program. Our suggestion, as part of Step 4 in the method, for factoring a

measure of the fidelity of program implementation into the program-wise effectiveness index, is not foolproof. In cases where program implementation is less than ideal, and there seems room for its improvement, our suggestion to inflate the program-wise effectiveness index according to the level of implementation makes sense. On the other hand, the implausible design of some programs will lead to their never being implemented according to some ideal; the actual (less than ideal) level of implementation observed may be as good as it can possibly get. In these cases, inflating the program-wise effectiveness index according to levels of fidelity of implementation does not make sense, and the fidelity of implementation measure (and its explanation) may better serve as a stand-alone index.

Third, for the example detailed in this paper, and in the synthesis method in general, we have resisted *a priori* ranking of the importance of the various pieces of evidence gathered around each SOS program goal. We believe that evaluators who are considering using the method will be savvy and flexible enough to recognize if and when, in their own evaluations, different sets of evidence should carry different *a priori* ranks and/or weights. In the current case, for example, we may justifiably have placed great stock in the statewide assessments of children's academic achievement, and thus weighted these data three times as heavy as any other piece of evidence in the analysis of academic achievement. In polling representatives of the various stakeholder groups, however, we found that while the Department of Education (our client) valued the statewide assessment findings considerably more highly than any other evidence, the opposite was also true for other stakeholder groups. Thus, in this case at least, we are not compelled to assign differential weights to evidence sets *a priori*, but rather are comfortable in letting the weightings emerge based on CoW ratings that provide gauges of the overall quality of each evidence set.

Lastly, this method is very much a work in progress. We know that we have not thought of everything, and that some of what we have thought of is less than perfect. As noted previously, we do not suggest that this is the final solution to the data synthesis challenges that evaluators face in mixed- or multi-method program evaluations. Rather, we see this method as a useful aide in assisting program evaluators reach and substantiate "bottom line" decisions. Further, our hope is that other evaluators will tailor the method for their own particular situations, try it out, and suggest improvements.

NOTES

1. M.Q. Patton suggested the "credibility" and "understandability" criteria to us via electronic correspondence, August 28, 2000.

ACKNOWLEDGMENTS

We gratefully acknowledge V.J. Caracelli, J.C. Greene, M.Q. Patton, D.L. Stufflebeam, A. Tashakkori for their thoughtful and highly-valued advice on "CoW" for assessing lines of evidence within multi- or mixed-method evaluations. The input of these evaluation luminaries challenged our thinking and strengthened our work. Any shortcomings or inaccuracies contained herein, however, are the responsibility of the authors alone. An earlier version of this paper was presented at *Evaluation 2000*, the annual meeting of the American Evaluation Association (AEA), Waikiki, Hawaii, November 1–4, 2000.

REFERENCES

Caracelli, V. J., & Greene, J. C. (1993). Data analysis strategies for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis, 15*(2), 195–207.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Denzin, N. K. (1978). The logic of naturalistic inquiry. In N. K. Denzin (Ed.), *Sociological methods: A sourcebook*. New York: McGraw-Hill.

Frechtling, J., & Sharp, L. (Eds.), (1997, August). *User-friendly handbook for mixed-method evaluations*. Arlington, VA: National Science Foundation.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*(10), 3–8.

Greene, J. C., & Caracelli, V. J. (1997). Defining and describing the paradigm issue in mixed-method evaluation. In J. C. Greene & V. J. Caracelli (Eds.), *Advances in mixed-method evaluation: The challenges and benefits of integrating diverse paradigms* (pp. 5–17). San Francisco, CA: Jossey-Bass. *New Directions for Evaluation, 74*.

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis, 11*(3), 255–274.

Hedges, L. V., Shymansky, J. A., & Woodworth, G. (1989). *A practical guide to modern methods of meta-analysis*. Washington, DC: National Science Teachers Association.

Hunter, J. E., & Schmidt, F. L. (1994). Correcting for sources of artificial variation across studies. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 323–336). New York: Russell Sage Foundation.

Jick, T. D. (1983). Mixing qualitative and quantitative methods: Triangulation in action. In J. Van Maanen (Ed.), *Qualitative methodology* (pp. 135–148). Beverly Hills, CA: Sage.

Joint Committee on Standards for Educational Evaluation (1994). *The program evaluation standards, how to assess evaluations of educational programs* (2nd ed.). Thousand Oaks, CA: Sage.

Kidder, L. H., & Fine, M. (1987). Qualitative and quantitative methods: When stories converge. In M. M. Mark & L. Shotland (Eds.), *Multiple methods in program evaluation* (pp. 57–75). San Francisco, CA: Jossey-Bass. *New Directions for Program Evaluation, 35*.

Kulik, J. A., & Kulik, C. C. (1989). The concept of meta-analysis. *International Journal of Educational Research, 13*, 227–234.

Mertens, D. M. (1998). *Research methods in education and psychology*. Thousand Oaks, CA: Sage.

Riggins, L., & Caracelli, V. J. (1994). Mixed-method collaboration. Mixed-method evaluation: Developing quality criteria through concept mapping. *Evaluation Practice, 15*(2), 139–152.

Patton, M. Q. (1987). *How to use qualitative methods in evaluation*. Newbury Park, CA: Sage.

Rossman, G. B., & Wilson, B. L. (1985). Numbers and words: Combining quantitative and qualitative methods in a single large-scale evaluation study. *Evaluation Review, 9*(5), 627–643.

Scriven, M. (1969). An introduction to meta-evaluation. *Educational Product Report, 2*(5, February), 36–38.

Scriven, M. (1971). Evaluating educational programs. In F. G. Caro (Ed.), *Readings in evaluation research* (pp. 49–53). New York: Russell Sage Foundation.

Scriven, M. (2000, June). *The logic and methodology of checklists*. Available online at http://www.wmich.edu/evalctr/checklists/logic_methodology.htm

Shotland, R. L., & Mark, M. M. (1987). Improving inferences from multiple methods. In M. M. Mark & L. Shotland (Eds.), *Multiple methods in program evaluation* (pp. 77–94). San Francisco, CA: Jossey-Bass. *New Directions for Program Evaluation, 35*.

Stufflebeam, D. L. (1974, December). *Meta-evaluation*. Paper #3, The Evaluation Center Occasional Paper Series. Kalamazoo, MI: The Evaluation Center, Western Michigan University.

Tashakkori, A., & Teddlie, C. (1998). *Mixed methodology: Combining qualitative and quantitative approaches*. Thousand Oaks, CA: Sage.

Trend, M. G. (1979). On the reconciliation of qualitative and quantitative analyses: A case study. In T. D. Cook & C. S. Reichardt (Eds.), *Qualitative and quantitative methods in evaluation research* (pp. 68–86). Beverly Hills, CA: Sage.

Waysman, M., & Savaya, R. (1997). Mixed-method evaluation: A case study. *Evaluation Practice*, *18*(3), 227–238.

Weiss, C. H. (1998). *Evaluation* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.