Educ Asse Eval Acc (2010) 22:5–28
DOI 10.1007/s11092-009-9089-8

Mixing methods in randomized controlled trials (RCTs): Validation, contextualization, triangulation, and control

James P. Spillane · Amber Stitziel Pareja · Lisa Dorner · Carol Barnes · Henry May · Jason Huff · Eric Camburn

Received: 3 December 2009 / Accepted: 22 December 2009 /
Published online: 14 January 2010

Abstract In this paper we described how we mixed research approaches in a Randomized Control Trial (RCT) of a school principal professional development program. Using examples from our study we illustrate how combining qualitative and quantitative data can address some key challenges from validating instruments and measures of mediator variables to examining how contextual factors interact

An earlier version of this paper was presented at the Annual Meeting of the American Educational Research Association, Chicago April 9th – 13th, 2007. The research was supported through a grant from the U.S. Department of Education's Institute of Education Sciences and the Distributed Leadership Studies supported by a grant from the National Science Foundation (Grant # EHR – 0412510). We also thank James Pustejovsky, Beth Sanders, and Jimmy Sebastian for their research assistance at various stages of the work. Please direct any correspondence regarding this paper to Jim Spillane – j-spillane@northwestern.edu.

J. P. Spillane (✉) · A. S. Pareja · L. Dorner
Northwestern University, 2120 Campus Drive, Annenberg Hall 208, Evanston, IL 60208, USA
e-mail: j-spillane@northwestern.edu

A. S. Pareja
e-mail: aspareja@northwestern.edu

L. Dorner
e-mail: dornerl@umsl.edu

C. Barnes
University of Michigan, Ann Arbor, MI, USA
e-mail: barnescl@umich.edu

H. May
University of Pennsylvania, Philadelphia, PA, USA
e-mail: hmay@gse.upenn.edu

J. Huff
Vanderbilt University, Nashville, TN, USA
e-mail: jason.huff@vanderbilt.edu

E. Camburn
University of Wisconsin-Madison, Madison, WI, USA
e-mail: ecamburn@education.wisc.edu

with the treatment. Describing how we transformed our qualitative and quantitative data, we consider how mixing methods enabled us to deal with the two core RCT challenges of random assignment and treatment control critical. Our account offers insights into ways of maximizing the potential of mixing research methods in RCTs.

Keywords Mixed methods · Randomized controlled trials

> "It is not enough to think well; we must also demonstrate the value and importance of a mixed methods way of thinking in our practice" (Greene 2006).
> "No commentator on evaluation devalues excellence with respect to experimental design, reproducibility, statistical rigor, etc. But we do say that these virtues are purchased at too high a price when they restrict an inquiry to what can be assessed with greatest certainty" (Cronbach 1988, p.7).

Some researchers have argued for more randomized controlled trials (RCTs) to be conducted in order to evaluate the efficacy of educational interventions (Boruch 2002; Cook 2002; Eisenhart and Towne 2003; Shavelson and Towne 2002). US policymakers have heeded the calls of RCT advocates; the Institute of Education Sciences (IES), for example, has devoted considerable funding to RCTs designed to determine the efficacy of educational programs (Shavelson and Towne 2002). During the years FY2002 through FY2004, for example, between 72% and 98% of the dollars awarded by the National Center for Education Research's Field Initiated Program were devoted to studies involving random assignment (Cook and Wong 2006).

While mixed method designs are relatively commonplace in evaluation research, they are rare in RCTs. Mixed method studies combine qualitative and quantitative research methods so they work in tandem to answer the key research questions in a single study (Johnson and Onwuegbuzie 2004; Yin 2006). Mixed method designs are increasingly popular in education and other applied fields (Chen 1997b; Mactavish and Schleien 2004; Nastasi and Schensul 2005; Sandelowski 1996). Some studies that claim to mix methods, however, are often parallel studies where qualitative and quantitative components are mostly independent of one another and weakly, if ever, connected systematically. Further, some scholars critique mixed method research, arguing that qualitative research is frequently assigned second-class status in such designs and its interpretive epistemological roots are undermined (Denzin and Lincoln 2005; Guba 1990; Howe 2004). Indeed, within the broader literature on RCTs, qualitative methods are treated primarily as a means for monitoring implementation or enhancing interpretation of quantitative results (Boruch 1997). Still, other scholars disagree, arguing that there is no reason that qualitative approaches need to be assigned a secondary role in mixed method designs (Creswell et al. 2006). These theoretical debates about mixed method designs, while important, may obscure how scholars are combining qualitative and quantitative research approaches in actual research studies (Maxwell and Loomis 2003).

Mixed method designs are relatively common in evaluation research, especially theory-driven evaluations (Chen 1990, 1997a, 2005; Gottlieb et al. 1992; Shavelson and Towne 2002; Weiss 1997). In this paper, we consider the role of mixed method research designs in a particular type of evaluation research design—the RCT. We

focus on RCTs because they face some particular challenges that can potentially be addressed in mixed method designs. Three core principles or assumptions in RCTs are *randomization*, *control*, and *comparison*; treatments are assigned *randomly* to subjects, treatments are *controlled,* and *comparison* of control and treatment groups enables us to detect a treatment effect, or the lack thereof. While randomization, control, and comparison are relatively easy to accomplish in laboratory settings, they are immensely more difficult in the real world where human and sociopolitical factors interact with assignments and treatments (Rossi et al. 2004). As a result of these interactions, we often end up with overlapping distributions between the treatment and control groups that undermine core assumptions of RCTs. A key goal of this paper is to draw attention to the potential power of mixed method research designs in RCTs by describing how we combined qualitative and quantitative approaches in an RCT that examined the impact of a principal professional development program (PDP) in one urban school district.

Our paper is organized in this manner: We begin by situating our work in the literature on mixed methods research, identifying various ways of combining qualitative and quantitative approaches. After describing the treatment, we overview the qualitative approaches we used in the RCT. Next, using examples from our study we illustrate how combining qualitative and quantitative data can address some key challenges from validating instruments and measures of mediator variables to examining how local contextual factors interact with the treatment. Next, we illustrate how we transformed some of our qualitative and quantitative data in order to combine different types of data. In doing so, we illustrate how mixing methods not only serves triangulation purposes but also addresses two key challenges in RCTs—random assignment and treatment control. We conclude with a discussion of how we can use qualitative and quantitative approaches in tandem in order to maximize the potential of mixed methods in RCTs.

1 Situating the work: Mixed methods in social science research and RCTs

While mixed method studies are increasingly popular in the social sciences, researchers often use the two approaches in parallel, rather than in tandem. As a result, the potential of mixing methods is not maximized. Still, over the past two decades, a body of work has emerged that either combines qualitative and quantitative approaches in a single research study or addresses the research design challenges in combining the two approaches (Creswell 2002; Morgan 1998; Morse 1991, 2003; Tashakkori and Teddlie 2003). In this section, we consider ways of mixing qualitative and quantitative approaches in social science research by describing some typologies that help structure the terrain. We then turn our attention to mixed methods in evaluation research and RCTs in particular.

There is no shortage of typologies for mixed methods research designs (Tashakkori and Teddlie 2003). Caracelli and Greene (1993) identify four mixed method data analysis strategies involving qualitative and quantitative data, some of them based on a review of evaluation studies. These include data transformation, typology development, extreme-case analysis, and data consolidation/merging. Data transformation involves translating one data type into another type in order to analyze both

together. Typology development refers to situations where analysis of one data type results in the development of a typology that is then used as the basis for analyzing another type of data. Extreme case analysis refers to situations where extreme cases are identified based on an analysis of one type of data and then these types are investigated based on analysis of another data type. The goal in this situation is to assess and enhance the original explanation for the extreme cases. Finally, data consolidation and merging involves the combined evaluation of both types of data to generate new or merged variables or data sets, which can be quantitatively or qualitatively defined and subjected to additional analysis.

Tashakkori and Teddlie (1998) develop a classification framework for mixed method research designs. They argue that one of the primary data analytic techniques used in mixed methods is the conversion of data gathered using one method into data from the other method in order to analyze the same data using alternative analytical approaches. This conversion can take place in two ways. First, transforming qualitative data into numerical codes that can then be analyzed quantitatively which they refer to as *quantitizing techniques* and *quantitized data* (see also, Miles and Huberman 1994; Sandelowski 2003). Second, transforming quantitative data into descriptions that can then be analyzed qualitatively referred to as *qualitizing techniques* and *qualitized data* (Tashakkori and Teddlie 1998, 2003).

Although parallel analysis of qualitative (QUAL) and quantitative (QUAN) data provides a richer knowledge of the variables and their associations, it is limiting since it only allows the researcher to use one kind of data analysis on each type of data (Tashakkori and Teddlie 1998). They argue that more information can be gleaned from the data through one of four approaches. The first approach involves using both qualitative and quantitative methods to simultaneously analyze the same data. The second approach is confirming and/or expanding the inferences generated from one method of data analysis (e.g., qualitative analysis) through a secondary analysis of the same data with another method of data analysis (e.g., quantitative analysis). The third approach consists of sequentially using the findings generated from one analytical approach (qualitative analysis) as the beginning point for an analysis of *other* data with the alternative approach (quantitative analysis). For example, researchers might classify individuals or events into groups based on a qualitative analysis of a data set and then with a new data set compare the prevalence of these groups in some population using quantitative approaches. The final approach is utilizing the results generated by one analytical approach (e.g., qualitative analysis of interview data) as the basis for collecting or analyzing new data using the other analytical approach.

Tashakkori and Teddlie (1998) develop a classification scheme of mixed method data analysis strategies that includes three major types. The first, *concurrent mixed analysis*, includes three types of analysis: 1) parallel mixed analysis, typically used for triangulation purposes; 2) concurrent analysis of the same qualitative data using both quantitative and qualitative techniques which necessitates quantitizing the qualitative data; and 3) concurrent analysis of the same quantitative data using both qualitative and quantitative techniques which involves qualitizing the quantitative data. Second, *sequential QUAL-QUAN analysis* involves an initial qualitative analysis that results in the identification of groups of individuals who are similar and then comparing these groups using quantitative techniques. This sequential

method can involve three approaches. The first consists of forming groups of people/settings/events based on qualitative analysis of qualitative data and then comparing these groups using quantitative techniques. The second approach is to identify sets of attributes or themes through qualitative analysis and then following this with confirmatory quantitative analysis. The third approach consists of using qualitative analytical techniques to establish a theoretical order of relations and/or causality and then using quantitative techniques to confirm the hypothesized relationship.

The third strategy for mixed method data analysis, s*equential QUAN-QUAL analysis*, involves quantitative analysis followed by qualitative analysis. Again, this method can involve three sub-approaches. The first approach is to form groups of people/settings/events based on quantitative analysis and then examining these groups using qualitative analytic techniques. The second approach consists of establishing categories of attributes or themes through initial quantitative analysis and then confirming these categories through qualitative analysis of qualitative data. The third approach is to explore quantitative data to find a theoretical order or relations and/or causality and then confirming the relations found with qualitative analysis of qualitative data. Both types of sequential analysis (QUAL-QUAN and QUAN-QUAL) may or may not involve qualitizing quantitative data or quantitizing qualitative data.

Using both qualitative and quantitative approaches is relatively common in evaluation studies (Caracelli and Greene 1993; Cook and Reichardt 1979; Datta 1994; Reichardt and Rallis 1994; Riggin 1997; Rossman and Wilson 1993). Evaluation studies employing mixed methods tend to involve either "component designs," where different approaches remain distinct and operate in parallel, or "integrated designs," where different approaches are combined so that they work in tandem (Caracelli and Greene 1997). One review of published articles involving mixed methods evaluations found that, of 57 articles considered, only five involved an integrative approach to analyzing the qualitative and quantitative data collected (Greene et al. 1989). For whatever reasons, component designs trumped integrated designs in evaluation studies that mixed qualitative and quantitative approaches. A more recent review of articles involving mixed methods, covering a ten year period starting in 1994 and using the Social Sciences Citation Index (SSCI), uncovered 232 articles in five fields: human, social and cultural geography; management and organizational behavior; media and cultural studies; sociology; and social psychology (Bryman 2006). Most striking, considering the focus of our paper, is the scarcity of experimental research designs. Fewer than fifteen studies involved experimental or quasi-experimental research designs (Bryman 2006). Another review, focusing on mixed methods in education research and not necessarily evaluation studies, found that of 1,156 articles reviewed from fifteen different journals, 145 involved qualitative and quantitative approaches (Niglas 2004). The available published work suggests that, with some exceptions, mixing methods in evaluation studies that involve either quasi-experimental trials (Cook et al. 2000; Lynch et al. 2007) or RCTs (Flemming et al. 2008; Hall and Howard 2008) is rare. This is surprising considering that evaluating the effects of programs in the field as distinct from in the laboratory is at the heart of RCTs and difficult to gauge without rich descriptions of contextual factors (Chatterji 2005).

2 Mixing methods in a RCT

We describe our RCT of a school principal development program in Cloverville, a mid-sized urban school district, in the southeastern US in this section.[1] Randomized experiments involving school principals are rare, with one recent review only identifying three such studies (Spillane et al. 2007). We begin with a brief description of the treatment and then turn our attention to the study design, detailing the qualitative and quantitative approaches we used in both data collection and data analysis. Describing how we actually mixed methods in data collection and data analysis, we identify three usages of mixed method designs in RCTs including monitoring the fidelity of treatment implementation, identifying variables that mediate relations between the treatment and outcomes, and validating measures and instruments.

2.1 The treatment

The principal development program was designed by an external provider to improve student achievement by developing principals' knowledge and skills for leading improvement in instruction. The program exemplified many of the characteristics associated with effective professional development. Among other things, participants had opportunities to work with particular topics over extended periods of time and apply their learning in their own work situation. In addition to workshops, the program involved study groups, case studies and action research projects, as well as distance learning experiences. Distributed over fourteen units, topics covered in the program included standards-based instructional systems, strategic thinking for principals, instructional leadership, effective student learning experiences, and developing a professional learning community.

Following a "train-the-trainer" model, faculty from the external provider organization that designed the program trained a leadership team from the district. The faculty then provided technical assistance when this local team subsequently trained the first cohort of Cloverville principals starting in summer 2005. While the local team was supposed to train a second cohort of Cloverville principals starting in summer 2006, this never happened due to changes in district leadership (see below). Further, principals in the early-treatment group were offered only half of the workshops in the program.

2.2 RCT design

Our RCT involved a delayed-treatment design in which half of the principals in Cloverville were randomly assigned to participate in the treatment at the beginning of the study (early-treatment group), and the remaining principals were randomly chosen to begin the treatment one year after the first group (delayed-treatment group). We excluded principals who were members of the Cloverville leadership team that would deliver the treatment to local school principals. To avoid the subversion of randomization (Boruch 1997), a research team member performed the

[1] Cloverville is a pseudonym.

random assignment. We used a basic random assignment design, incorporating school level as a blocking variable and checking the randomization process by comparing early and late-treatment principals on a range of variables. We checked the randomization process by comparing the two groups of principals on variables measuring both school and principal characteristics including gender, race, years of experience and whether the school had met Adequate Yearly Progress (AYP). We found that the two groups were almost identical on each variable.

The appointment of a new school district superintendent mid-way through the first year of the study (fall 2005), however, meant that we had to change our original research design shortly after it began. Specifically, the new school district leadership decided not to give the professional development program to the principals in the delayed-treatment group so our RCT became a straightforward randomized trial. As we will discuss below, these changes in district leadership also influenced the participation of some principals in the early-treatment group in the workshops.

We used a theory-driven evaluation in order to facilitate strong causal inferences on efficacy and enable contributions to basic social theory (Birckmayer and Weiss 2000; Chen and Rossi 1980; Lipsey and Wilson 1993). Theory-driven evaluation involves using the substantive theories about the relationships between a program's treatment variables and outcome variables to guide the design of the RCT (Chen and Rossi 1983; Shadish et al. 1991). Our logic model (see Fig. 1) posits that school principals will acquire new knowledge and skills through participation in the principal development program; but what principals learn will depend on both the content and pedagogy of the workshops they attend *and* their background. In turn, principals' new knowledge and skill will contribute to change in school leadership practice, effecting change in those school conditions (e.g., trust, collective responsibility, academic press, etc.) that are believed to be critical for improving classroom teaching. Improvements in classroom teaching will in turn lead to improvements in student achievement.

In our logic model, principal knowledge and practice are proximal outcomes and student achievement is a distal outcome. At the same time, principal knowledge is a mediating variable between the treatment and principals' practice. Further, relations

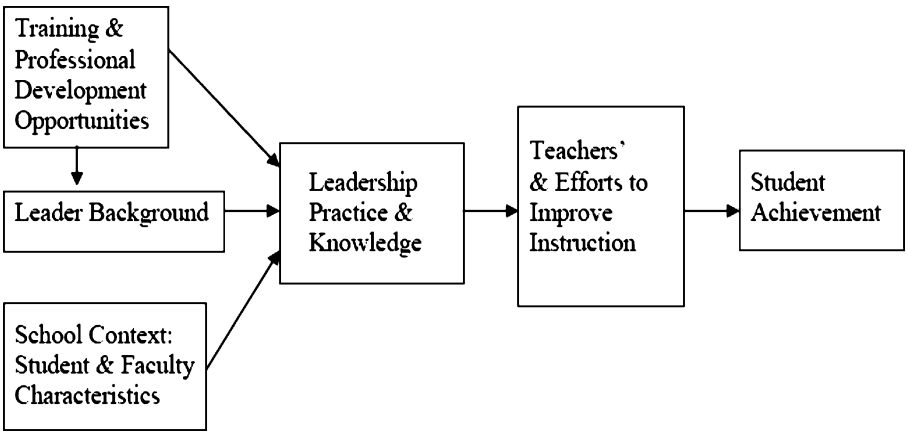

Fig. 1 Conceptual framework

between the treatment and student achievement will be *mediated* not only by principals' knowledge and practice but also by school level conditions and classroom teaching. We acknowledge that the relations among these variables are likely bi-directional. Mediator variables are important to understanding and testing the intervention's theory of action because it is through transforming the conditions that these variables measure that an intervention has an effect (Petrosino 2000; Weiss 1997).

2.3 Data collection overview

We discuss the particulars of data collection and data analysis below as we describe how we combined qualitative and quantitative approaches in our RCT. In this section we provide a brief overview of our data collection approaches, study timeline, and the sample. Beginning in spring 2005, school principals in both the early-treatment and late-treatment groups and the staff in their schools were studied over three school years. Data collected in spring 2005 prior to the beginning of the professional development program served as baseline data for the RCT. Because participation in the study was voluntary, we provided financial incentives to study participants based on their completion of different research instruments (e.g., questionnaires, logs, etc.) (Boruch 1997). Principals who completed all research instruments received incentives amounting to $235 annually.

From the outset, we mixed quantitative and qualitative approaches in various ways and to various ends. Quantitative approaches included a school staff questionnaire (SSQ) and school principal questionnaire (PQ), principal End-of-Day (EOD) and Experience Sampling Method (ESM) logs, structured observations, and student achievement data that was provided by the school district. The PQ was administered annually online starting in spring 2005 for three years. The response rates for 2005 and 2006 were 94% and 77% respectively. The SSQ was administered via mail in 2005 and again in 2007 to nearly all staff in every school in the district with a response rate of 87% in 2005 and 78% in 2007. Principals were asked to complete ESM and EOD logs for six consecutive school days in spring 2005. They completed the EOD log again over the course of five consecutive days during six subsequent periods: fall of 2005, winter of 2006, spring of 2006, fall of 2006, winter of 2007, and spring of 2007. Response rates ranged from a high of 93% (spring 2005) to a low of 67% (spring 2007).

Qualitative approaches included observations of the program delivery (i.e., professional development workshops) followed by post observation interviews, observations of a sub-sample of school principals' practice over the course of an entire workday followed by in-depth cognitive interviews, principals' responses to open-ended scenarios, and interviews with district office staff. For example, we conducted 63 interviews with 20 principals, including at least one interview with every participant who attended a workshop session.

2.4 Program effects

Though not the primary focus of this paper, our RCT was designed to evaluate the effects of a principal professional development program. With respect to gains in student achievement, our results were not statistically significant in either the 'intent

to treat' (IT) or 'treatment on the treated' (TOT) analyses. The only significant impact detected was in our analyses of the treatment on those who received it. In those analyses, we found that principals who in fact participated in the treatment spent significantly more time on planning and goal setting in practice as measured by the EOD than did non-participants. The lack of statistically significant results may indicate that the professional development program did not have a beneficial effect on Cloverville principals. Another possibility is that the results presented reflect the program in its early stages, and that over time effects of the program will emerge. Our experiences in conducting the RCT, however, illustrate how mixing methods can contribute to the returns in terms of knowledge generated by RCTs.

3 Combining qualitative and quantitative data in a RCT: Validation & contextualization

Sometimes mixing methods is relatively straightforward in that it involves combining data generated by different research approaches, without any fundamental data transformation, in order to gain new insights into an issue or variable. To illustrate how we mixed methods, we discuss below our efforts to measure two key mediator variables in our logic model (Fig. 1)—*principal knowledge* and *principal practice*. To measure changes in *principal knowledge* (proximal outcome) we combined quantitative and qualitative approaches. Items on the PQ and the SSQ measured principal knowledge for both the early-treatment and delayed-treatment groups. On the PQ, we adapted a version of *The School Leadership Self Inventory* (National Policy Board for Educational Administration 2002), a self-reporting inventory consisting of Likert scale items based on the Interstate School Leaders Licensure Consortium (ISLLC) standards. These items read: "This question asks about your knowledge in a variety of areas of school leadership. For each area please indicate the degree to which you believe your current knowledge reflects personal mastery (knowledge and understanding of the area)." The stem then read, "To what extent do you currently have personal mastery (knowledge and understanding) of the following:" and the choices were a 5-point scale: a little, some, sufficient, quite a bit, and a great deal. We used these items to tap into principals' perceived expertise about standards-based reform, data-based decision-making, and principles of teaching and learning. On the SSQ, we investigated principals' knowledge by measuring teachers' perceptions of their principal's understanding of the principles of effective teaching and learning. We used a three item scale (alpha=.92) with questions such as: "Please mark the extent to which you disagree or agree which each of the following: 'The principal at this school has a strong understanding of…'" followed by another scale of answers.

We combined our quantitative approaches to gathering data on principals' knowledge with qualitative approaches that involved open-ended scenarios, observations, and cognitive interviews. We asked all principals to respond to a video simulation and five written scenarios in spring 2005 (pre-treatment) and again in spring 2007 (post-treatment). The video simulation used footage of a teacher teaching whereas the five written scenarios, varying in length from 68 to 145 words, described

brief, school-related problems. Principals were given 45 min to write open-ended narrative responses to the problems posed in the scenarios. Forty-six principals responded to the scenarios in spring 2005 and 43 principals responded in spring 2007. The average number of words written per scenario was 84.8, ranging from 115.7 for scenario 1 (video simulation) to 71.9 for scenario 6. Length of response was not correlated with the placement of the scenario—response to prompt 2 of the simulation, which came first, generated the shortest response with an average word count of 63.7. As mentioned above, we also conducted 63 semi-structured post-observation interviews with 20 principals; these were typically held either after a professional development workshop or at the end of a day spent observing principals at work; the timing of these interviews ensured that much of the content was grounded in particular events. Among other things, we used these interviews to investigate principals' professional learning and how they used knowledge in their practice.

We also combined quantitative and qualitative approaches in examining changes in *principal practice*. To quantitatively investigate changes in principal practice we used the PQ, the SSQ, the End-of-Day (EOD) log and the Experience Sampling Method (ESM) log. To investigate qualitative changes in principals' practice we used a combination of structured and semi-structured observations of the program workshops and school principals at work together with post-observation interviews. Based on the PQ data, we constructed two measures of principal practice, one focusing on principals' involvement in planning/goal setting and the other focusing on their monitoring of instruction. Our measure of planning/goal setting (alpha=.89) was made-up of five items on which principals reported the frequency with which they set timelines for improvement, worked on plans to improve teaching, clarified expectations in the school, and framed and communicated improvement goals. The monitoring instruction measure (alpha=.83) consisted of four items capturing the frequency with which principals observed teachers and monitored instruction and curriculum implementation. Both sets of items had these response choices: 1=never, 2=a few times throughout the year, 3=a few times per month, 4=1–2 days per week, and 5=more than 2 days per week. On the SSQ, we also included questions to measure teachers' perceptions of their principals' practices as distinct from knowledge. The stem for these items was: "Please mark the extent to which you disagree or agree which each of the following: The principal at this school…" and then listed particular practices. For example, one item asked teachers to report on the extent to which principals monitor instructional improvement (five item scale, alpha=.85).

Principals completed the End-of-Day (EOD) logs online at the end of each school day over the course of number of consecutive days at different times during the year. We used the EOD to gather data about principals' daily practice. Specifically, principals reported how much time *during each hour of the day* between 6 a.m. and 7 p.m. they spent participating in nine types of activities: building/operations, finances, community/parent, school district, student affairs, personnel, planning/setting goals, instructional leadership, and professional growth. Principals completed the EOD log over the course of seven different periods that ranged from five to six consecutive days in duration with response rates ranging from a high of 93% to a low of 67%. With the ESM, used in the spring of 2005, we paged principals

randomly via palm pilots approximately 15 times a day. At each beep, principals completed a short questionnaire in which they reported on what they were doing and how they were doing it.

To validate some of our research instruments and measures of practice, we combined different types of quantitative data. For example, using data generated by the ESM log in tandem with data generated by the EOD log, we were able to validate the EOD log. Specifically, to assess the accuracy of the daily log we compared results from the EOD log to results from the ESM log for the same time period. The validity of the EOD log was assessed by comparing estimates of the percentage of time principals spent on six domains of principal practice measured by the EOD and the ESM logs—building operations, finances, student affairs, personnel issues, instructional leadership, and professional growth. Overall, the EOD and ESM yielded similar estimates of the frequency with which principals engage in the six functions, generating identical estimates of the frequency of two functions (e.g., dealing with personnel and professional growth) and rank ordering the six functions almost identically. While the estimates for building operations, finances, and student affairs produced by the ESM and the EOD differed more, these differences were still less than 5 percentage points. We return to the issue of log validity below in the sub-section on data analysis.

To investigate school principals' practice (as well as their knowledge in use), we used semi-structured and structured observations of principals' practice as well as post-observation cognitive interviews. These qualitative approaches were intended to complement and enrich the quantitative measures generated through the EOD, ESM, and PQ and SSQ. We shadowed several principals for one EOD logging day in February of 2006 and again in February of 2007. In 2006, we shadowed a total of 15 principals, twelve from the early-treatment group and three from the delayed-treatment group. In 2007, we shadowed 13 principals, eight from the early-treatment group and five from the late-treatment group. These observations included both a structured (quantitative) and semi-structured (qualitative) component. First, when paged at 15-minute intervals, researchers documented principals' practice using a standardized observation guide aligned to the EOD and ESM logging categories (e.g., where the principal was, the type of activity) and then provided a written description of what the principal was doing. Second, between each 15-minute interval researchers wrote thick descriptions of what the principal was doing.

Each observation was followed by cognitive interviews, 45 and 60 minutes long, in which interviewers used a cognitive explanation protocol (Chi 1997) to prompt principals to recall prior, practice-based cognitive performances from a recent "naturalistic" context (Klein et al. 1989). Specifically, each principal was asked to describe two activities s/he participated in that day, including a specific instance of instructional leadership aligned to our logging categories. Principals were also asked to describe an instance in which they used knowledge from the treatment program in their daily practice. For each situational prompt, interviewers also asked principals how and why they logged the activity in the EOD log.

Combining qualitative and quantitative approaches to gather data on principals' practice enabled us to do two things. First, we were able to validate the EOD and ESM logs using our observations of a sub-sample of school principals. A key concern with any research instrument is whether it captures the phenomenon it is

designed to measure. We "shadowed" a subset of five randomly selected principals during the spring of 2005 as they completed the EOD and ESM logs, writing narrative reports of that on-site shadowing visit (OSV). Every ten minutes, the researcher recorded the activity in which the principal was engaged, along with a description of the context in which the activity occurred. In addition, when the principal was alerted (paged) to complete the ESM instrument, the researcher also responded to a subset of the ESM questions. Comparing data generated by the ESM log and OSV data, for example, we computed the associations between the ESM and the observer data and determined whether these associations were important and significant. Our concurrent mixed analysis found significant and high agreement between the ESM data and observer data on whether the principal was leading the activity, and whether the principal was co-leading with a teacher or non-teaching staff person. In addition to serving validation purposes this sort of concurrent mixed analysis also served triangulation purposes.

Second, we were able to investigate reasons as to why principals might not record particular activities in their EOD log, generating knowledge for redesigning the log. As described above, our comparison of quantitative EOD log and ESM log data found that while overall agreement was high, principals under-reported building operation and finance-type activities in the EOD log compared with the ESM log. To explore these findings, we analyzed qualitative field note data from our shadowing of the five school principals in spring 2005. Focusing on the finance- and building operation-type activities, we identified 20 instances in the field note data where the principal failed to report an activity in the EOD log but the observer did report that activity for that particular hour. This analysis involved a QUAN-QUAL sequence. Focusing on field note data for two of the five principals, we generated and defined a series of working hypotheses developing a typology of possible reasons for why principals might fail to report an activity.

Based on our analysis of these data, we developed four working hypotheses as to why school principals might fail to log activities. First, the brevity hypothesis states that when an activity is brief and scarce in a particular hour it is less likely to be logged by the principal. Second, the non-continuous hypothesis refers to whether an activity is continuous over some time segment and un-interrupted or blended with other activity types. It can be thought of at the level of any one-hour period (the unit of recording for the EOD log) where continuity refers to the fact that the type of activity spanned two or more consecutive 10-minute segments. Second, it can be thought about at the level of any 10-minute segment (the unit of recording for the observer's shadow data) and refers to the fact that this is the only activity for that 10-minute segment. Third, we hypothesize that activities that take place early in the hour may be more easily recalled than activities that take place in the middle of an hour—the sequencing hypothesis. Fourth, the regularity hypothesis refers to whether an activity happens regularly.

We then qualitatively analyzed the shadow data for all five principals to see if the four hypotheses were tenable and worth further consideration. Based on our qualitative analysis of the observation data from the five principals, we refined and more clearly specified the brevity hypothesis and articulated a new hypothesis, the overshadowing hypothesis, which states that activities that occur in the same hour block as more dramatic or significant events are less likely to be logged. As we will

discuss in the next section, another step in this work involved *quantitizing* the qualitative field note data.

Experimentation in the real world is susceptible to changes in the social and political environment (Rossi et al. 2004). Hence, close attention to the local context *and* the manner in which it interacts with the treatment is critical in RCTs. Quantitative data (e.g., data on principal attendance at professional development workshops) combined with qualitative data (e.g., interviews with district office staff and school principals) enabled us to understand how changes in the local environment influenced the treatment delivery. Quantitative attendance data at the professional development workshops suggested a problem with the delivery of the treatment. There was evidence of the subversion of treatment assignment and non-participation in attendance data from the first workshop in June 2005. First, only 12 of the 24 principals assigned to the early-treatment group attended the first workshop. Second, three principals who were assigned to the delayed-treatment group attended the workshops for principals in the early treatment group. Third, during the 2005–06 school year, the three principals from the delayed-treatment group were regular attendees at the 5 professional development workshops and substantial numbers of principals assigned to the early-treatment group were no shows at these sessions.

Combining the quantitative data on attendance with interview data from both district office staff and school principals provided insights into how shifting local conditions subverted the delivery of the treatment. The forced retirement of the Cloverville superintendent, who was responsible for bringing the principal professional development program to the district, and the arrival of a new superintendent in fall 2005 contributed to shifts in district office priorities. The new superintendent brought his own ideas and preferences for school principal development to Cloverville. Further, there were changes in senior district office staff, namely the departure of the district's director of professional development during the 2005–06 school year. Thus, the new superintendent's other professional development program competed with the existing program for school principals' attention. These changes resulted not only in the treatment being cancelled for the principals in the late-treatment group but also influenced the participation of principals in the early-treatment group. Our analysis of principal interview data suggested that shifts in the district's priorities were not lost on principals in the early-treatment group, influencing their participation in and engagement with the treatment. Our mixed method approach enabled us to examine the interaction between this situation and the treatment in detail.

To summarize tentatively, mixing qualitative and quantitative methods allowed us to do a number of things in our RCT. First, we were able to triangulate findings generated by different data sources on principal knowledge and principal practice, two core mediator variables. Second, we were able to validate our log instruments and our measures of some core constructs such as principal knowledge. Third, we were able to identify possible reasons as to why certain events were not being logged by principals—information that was important for redesigning our log instruments. Fourth, we were able to use quantitative data to purposively sample principals for qualitative data collection, both through interviews and observation. Fifth, we were able to develop rich understandings of the local context and, more importantly, how

aspects of the local context interacted with the treatment to shape the ultimate delivery of the treatment. While scholars may be able to anticipate and potentially avoid some of the threats to their interventions posed by the local context by seeking out school systems that are supportive of their treatments (Borman et al. 2005; Rossi et al. 2004), our account suggests that local support for an intervention can change suddenly.

4 Transforming qualitative and quantitative data: Triangulation and treatment control

Our account thus far has focused on combining data generated from qualitative and quantitative approaches for validation, triangulation, and research instrument redesign purposes. In an effort to maximize the returns from our mixed method RCT design, we also transformed some data in our analyses by *quantitizing* qualitative data and *qualitizing* quantitative data (Tashakkori and Teddlie 2003). For example, we quantitized field note data from our observations of school principals and then combined these data with other quantitative data (e.g., log data, PQ data) in order to conduct a preliminary test of our working hypotheses (described above) as to why principals do not record some activities in the EOD log. Three researchers independently coded for each instance of the four hypotheses for the five school principals. Calculating inter-rater reliability, we found strong agreement between coders for the brevity (.8), non-continuous (.8), and overshadowing (.85) hypotheses, but agreement rates for the sequencing hypothesis and regularity hypothesis were low. We then engaged in a reconciliation process whereby we reviewed each individual case on which there was disagreement which in turn resulted in more specification of the sequencing hypothesis and dropping the regularity hypothesis. This QUAN-QUAL-QUAN sequence of analysis, involving the quantitizing of qualitative field note data, is the basis for a fourth step (i.e., QUAN). Specifically, using field note data on sixteen school principals observed in the second year of the study (spring 2006) together with their EOD log data for the same day, we plan to test our hypotheses as to why principals may fail to log certain activities. Hence, we have a QUAN-QUAL-QUAN-QUAN sequence of mixed methods analysis.

While the transforming of data enabled us to address issues of validity, it also enabled us to address two other issues that are unique to RCTs. Two key assumptions in RCTs are the random assignment of subjects to the treatment and the control of the treatment. Hence, the professional development program was *randomly* assigned to half of Cloverville's school principals (early-treatment group). Further, the plan was for the professional development to be controlled so that only those principals assigned to the early-treatment group would receive the treatment in the first year of the study. While these designs are standard in RCTs, the reality is that treatments are rarely as well controlled as they are in laboratory situations or indeed in standard medical trials. Human and social factors in the situations in which educational treatments are deployed can complicate evaluation because they often threaten assumptions about randomization and control (Rossi et al. 2004). Treatments, like the professional development program in our RCT, interact with local conditions. Indeed, the control of treatments in the wild will depend to some

degree on the endorsement of local school system actors *and* on how those being treated take to the treatment. As a result, in RCTs we often have overlapping distributions of treatments between the treatment group and the control group. For example, some individuals assigned to the treatment group end up not attending or attending infrequently.

We combined qualitized quantitative data and quantitized qualitative data with qualitative and quantitative data in order to better understand changes in school principals' knowledge, a proximal outcome and key mediator variable in our RCT (see Fig. 1). Finding no evidence of a program effect on student achievement (distal outcome) from both an 'intent to treat' and 'treatment on the treated' analyses, we decided to examine changes in two proximal outcomes—principals' knowledge and practice—also key mediator variables. We focus chiefly on principals' knowledge in this section to illustrate how we mixed methods in our data analysis in order to examine changes in principals' knowledge over time. This work not only involved combining different data types but also necessitated the transformation of data (i.e., qualitizing and quantizing). Focusing on principal knowledge, we elaborate on the following three stages in this analysis below:

- Quantitizing qualitative scenario data
- Combining quantitized scenario data with quantitative PQ and SSQ data
- Qualitizing quantitative PQ, SSQ, and attendance data and combining it with qualitative scenario, observation, and interview data

We describe each step to show how these analyses complicated what constituted our RCT's treatment group, as there were different and *overlapping* treatment groups.

In the first stage we quantitized the qualitative scenario data collected in 2005 (pre-treatment) and again in 2007 (post-treatment). Specifically, we developed a set of rubrics corresponding to a five-point scale (see Appendix for example) for principals' understandings of core competencies including effective teaching and learning, standards based reform, and data-based decision-making. Our competency rubrics were aligned with the PQ competency items and the treatment curriculum. Two coders worked independently and used the rubrics to score principals' responses to each of the six open-ended scenarios. After calculating inter-rater reliabilities, we worked to adjudicate any disagreements between the ratings of the two coders through an arbitration process that involved one of the coders and another researcher. Based on this arbitration process, a final score was assigned to each scenario for each principal.

A second stage in this analytic work involved combining our quantitized scenario data with our quantitative SSQ, PQ, and attendance data in order to analyze changes in early-treatment and delayed-treatment principals' knowledge and practice over time (i.e., pre- and post-treatment). Before using these data to examine change in principals' knowledge, however, we explored relations among our three sources of data on principal knowledge in order to triangulate across different data sources. Comparing principals' quantitized scenario scores with their self-reports of their knowledge on the PQ and teachers' ratings of principal knowledge on the SSQ, we found no meaningful or significant correlations. For example, with respect to principles of effective teaching and learning, the correlation between the PQ self-report measure and the scenario measure was only .04. Similarly, the correlations

between teachers' ratings of principal expertise (SSQ) and principals' self-reports (PQ) for principles of effective teaching and learning were only .27. In contrast, correlations between teachers' ratings of principals' knowledge on the SSQ were more highly correlated with the principals' scenario scores with a correlation of .43 for principles of effective teaching and learning (Goldring et al. 2009).

These efforts at triangulation raised several issues about our study operations and measures of principal knowledge that are the subject of ongoing work. First, principals' perception of their expertise was not related to the level of expertise demonstrated in their scenario responses. There is some evidence to suggest that people in general are not good at assessing their own expertise in a domain—incompetent people don't know that they are incompetent and competent individuals tend to underestimate their competence (Kruger and Dunning 1999). Further, the scenarios and PQ items may be measuring different constructs. Second, the higher correlations between teachers' ratings of their principals' expertise and the scenarios suggests that asking teachers about their principals' knowledge may be a more accurate way of tapping into that knowledge. Of course, in interpreting these relationships we must recognize possible two-way relationships between the measures. Overall, the divergent findings from our triangulation efforts pressed us to re-examine our study operations and measures for principal knowledge. While convergence across data sources is important for triangulation, divergence can also play an important role in understanding the phenomenon under study.

We also tested for change in principals' knowledge by comparing treatment and control principals' gain scores on their scenario responses, gain scores from their self-reports on the PQ, and gain scores from teachers' reports on principals' knowledge. Our analysis here focused on 24 of Cloverville's 52 school principals, twelve who originally were randomly assigned to the treatment and twelve who were randomly assigned to the control group. Three of those assigned to the control group crossed over into the treatment group at the first workshop. Our analyses across the three different types of quantitative data found few significant differences in principals' mean gain scores between the treatment and control groups by 2007. These analyses suggested rather sobering treatment effects; principals who were "treated" did not develop significantly more knowledge than their colleagues in the control group. This lack of significant difference is, in part, a function of the small number of participating principals (n=24) which makes it less likely that results will reach statistical significance. It is also likely that shifts in district office priorities, as we discuss above, also mattered here. Still, shifting gears somewhat, we zeroed in on principals in the treatment group, adopting a qualitative approach to both our qualitative and quantitative data in an effort to understand how these principals engaged the treatment and how their learning, knowledge and practice evolved over the course of the study.

A third stage in our analysis then involved a "data consolidation" technique (Tashakkori and Teddlie 1998) to develop case studies based upon both qualitative and quantitative data sources. This data consolidation involved the combined evaluation of various types of data to generate new data sets that can be subjected to additional analyses. In order to do this we qualitized the longitudinal quantitative data generated by the PQ, SSQ, and logs. We then combined these qualitized data with our qualitative scenario, interview and observation data. Drawing from these different data sources we explored, among other things, participants': (1) attendance

and level of engagement in the professional development workshops; (2) motivation to learn from the workshops as captured in their talk about the workshops and their discussion of implementing ideas gleaned from the workshops; (3) experiences and other professional learning activities; and (4) school and career situation.

We also qualitatively coded principals' responses to the scenarios at the two time points (i.e., pre- and post- treatment) following a grounded theory methodology (Strauss and Corbin 1990). Using the general guiding question, "How, if at all, do these scenario responses exhibit expertise in leadership?" we first openly coded each pre- and post-scenario response. Our initial analyses were "blind" to which principals were treatment versus control to ensure that we did not unconsciously look for change only in the principals who attended the PDP; we later returned to the data knowing who participated and who did not. We found that the responses varied in how much principals suggested (1) exploring a problem through further research; (2) implementing a solution, with or without further research; and (3) considering "if-then" scenarios of possible, contingent solutions. Re-reading principals' scenario responses, we then considered whether and how their answers showed development in both problem-solving expertise as well as content knowledge (as shown through their use of concepts taught in the program lessons).

Based on these analyses we constructed cases that developed a more nuanced understanding of changes in principal knowledge over time and the factors that might account for these changes or the lack thereof. Overall, our examination of cases followed a qualitative "set theoretic" data analysis procedure (Ragin 2000) in which we examined the overlap of the different cases on different dimensions that might account for changes in their knowledge. Our cases fell into one of four groups: (1) engaged and enthusiastic principals who had varying years of experience and who showed qualitative growth in their scenario responses in certain domains; (2) less enthusiastic but somewhat engaged, novice principals, some of whom gained some expertise; (3) more experienced principals who were not engaged in the PDP, but were enthusiastic about other learning activities; or (4) late-career principals who dropped out of the workshops entirely and showed little expertise development over time. It is important to point out that these groupings of principals are *not* simply based on the number of professional development workshops they attended—*treatment dosage*. For example, two of the four principals in Group Two attended all of the workshops whereas two of the five principals in Group One did *not* attend all of the workshops. Comparing these different groups of principals over time, we concluded that some principals who participated in the treatment did develop new knowledge by 2007 and we theorized how personal and situational factors help account for differences among the groups.

Our analysis suggests that what counted as being treated in our RCT differed depending on the particular principal. Rather than having two distinct groupings (i.e., treatment and control), we had multiple and at times overlapping distributions of treatments and control principals. At one level, because three principals from the control group switched into the treatment group—"crossovers" (Bloom 2005)—we ended up with our treatment and control groups overlapping. At another level, within the 'revised' treatment group, attendance differed dramatically with some principals attending no workshops—"no-shows" (Bloom 2005)—while others attended all of the workshops. Our qualitative analysis of qualitative and quantitative data pressed even further on what it meant to be treated in our RCT (Boruch 1997). While we

took into account principals' attendance at workshops, we also examined, among other things, principals' engagement with the material presented and their use of ideas gleaned from the workshops in practice. Based on this analysis, we identified four different 'levels' of treatment among the principals in the 'revised' treatment group and found that principals in some of these groups did change their knowledge over the course of their study. By mixing methods we were able to show that the treatment was not well controlled with multiple and overlapping distributions of what it meant to receive the treatment.

5 Discussion and conclusion

Evaluating interventions designed to influence complex social phenomena, such as principal knowledge and practice, in the real world presents new challenges for collecting data and verifying inferences from these data. Based on our work, we argue that researchers can benefit from employing correspondingly complex research designs and analytic strategies—designs and methods that provide as much evidence as possible so that the informed reader can agree or disagree with the conclusions drawn by the researchers (Cronbach 1988; Messick 1988). Further, we believe these designs benefit from employing interpretive as well as psychometric or survey methods (Moss 1994). Given the emerging press for RCTs in evaluation studies of educational interventions, using mixed methods for collecting and analyzing data not only adds rigor to the conclusions produced by such studies but also can increase the returns from such evaluations in the form of knowledge related to instrument and measure validation and core constructs. Data from multiple sources, both qualitative and quantitative, allowed us to unpack core constructs such as principals' knowledge and practice and also informed us of possible threats to validity. While validation, triangulation and implementation fidelity are not challenges that are unique to RCTs, they are critical considerations in such studies that, as we demonstrate, can be addressed through mixed method designs. Further, our account shows how mixing methods can contribute to better understandings of the challenges of contextualization and treatment control in RCTs.

In the RCT described in this study, both qualitative and quantitative approaches were used *together* and *in tandem*. From the outset, project researchers never treated the qualitative approaches to data collection in our RCT as somehow secondary to the quantitative approaches. Moreover, during the data analysis phase we continued to mix qualitative and quantitative data and analytic approaches. In part, this openness to mixing methods reflected the composition of the research team, which included scholars who were chiefly qualitative, primarily quantitative, and some who mixed methods. The balancing of qualitative and quantitative approaches was more than likely also aided by the absence of a main effect in our quantitative analyses, pressing the research team to dig deeper in order to discern the effects of the treatment on the ground. Regardless, our account offers existence proof that in mixing methods, even in RCTs, qualitative approaches do not have to play secondary or supporting roles. It also merits noting that a core part of our work, though not the primary focus of this paper, involved combining different types of quantitative approaches and mixing different types of qualitative approaches.

By mixing methods in our RCT we generated both convergent and divergent findings. Some quantitative methods or analyses produced findings that conflicted with our qualitative approaches. It is often tempting to ignore divergent findings as they can initially be seen as undermining efforts to triangulate particular findings. In our experience, this is a mistake. So we urge caution, encouraging engagement with divergent findings rather than reverting to a quick-fix consensus. In our work, divergent findings contributed to our interpretations of the data, helping to generate new questions and suggesting new lines of analysis. Examining quantitative findings *with* qualitative data, even when they diverged, pressed us to consider alternative explanations and to pursue new analyses in order to better understand the patterns we were finding. At the least, divergent findings between quantitative and qualitative data create puzzles for researchers that are sometimes more informative than "convergent" findings. Solving or even addressing such puzzles can infuse research with more rigor and findings with more authority (Cronbach 1989; Denzin 1978, 1989; Mathison 1988; Moss 1992, 1994). One potential pitfall here is following up on every divergent finding rather than strategically and selecting those divergent findings that are likely to generate the greatest returns.

Considering that some RCTs find no evidence of a treatment effect, mixed method designs increase the probability that such studies will generate other valuable empirical knowledge in addition to evidence of the absence of a treatment effect. Our efforts to validate research instruments and to measure mediator variables and proximal outcomes, for example, generated knowledge that can be used in the redesign of the intervention and/or the development of new interventions. Further, this knowledge can also be used in developing new and improved research instruments, study operations and measures for principal knowledge and practice, among other phenomena.

Because most RCTs are probabilistic, showing effects 'on average', the work often has limited utility in guiding policy-makers' and practitioners' work in particular situations. While an efficacy study might show that a particular intervention works or does not work on average, it offers limited practical knowledge essential for bridging the research-policy/practice divide. Mixed method designs can help in this situation as they generate knowledge that is conducive for translating research for the world of practice. Though our RCT did not find evidence that the intervention worked, by mixing methods we were able to identify circumstances under which *some* principals did change their knowledge and practice, suggesting practical knowledge about what it might take for a principal professional development program to have an impact. Our mixed methods analyses helped us to pinpoint what was happening for those principals who were treated, including how they understood and engaged the workshops, and how these factors influenced what it meant to be treated.

Mixing methods also necessitates not losing touch with the particular ontological or epistemological fundamentals of either qualitative or quantitative research. While quantitative research often assumes a single underlying truth, qualitative research finds its roots in an interpretive framework that allows for multiple ways of understanding the same phenomenon. Certainly both traditions have things in common and in difference. While we agree that quantitative and qualitative forms of research can be compatible (Brewer and Hunter 1989; Howe 1988; Reichardt and

Rallis 1994), the challenges of mixing these approaches should not be understated. We met this challenge in part through a multi-method research team. In this paper, we aimed to shake up notions of a strict and vast divide between "qualitative" and "quantitative" approaches, even in RCTs. We realize that our message may be a difficult sell in an increasingly polarized research and policy environment, especially in the US, where quantitative approaches are increasingly in vogue. Based on our work, we realized the potential for understanding to be found in the interaction of the two traditions. Fundamentally, the importance lies not with what kind of data we collect, but with using multiple perspectives to recognize how to approach these data and what may be gained and lost. Still, it is important that we subject our qualitative data to qualitative analytical approaches and that we not get carried away with applying *quantitizing techniques* to our qualitative data. In quantitizing qualitative data it is easy to lose touch with study participants' understandings of their worlds. The same holds true for qualitizing quantitative approaches. Ultimately, the key is to pinpoint what it is we want to know and subsequently understand and acknowledge that there are different things to know and multiple ways of knowing.

Appendix

Effective teaching and learning scenario coding rubric

Dimensions of teaching and learning referred to in the scale below include but are NOT limited to:

- student and/or teacher effort produces achievement,
- student learning is about making connections,
- students learn with and through others,
- student learning takes time,
- student and teacher motivation is important to effective teaching and student learning,
- focused teaching promotes accelerated learning,
- clear expectations and continuous feedback to students and/or teachers activate student learning (this does not include the process of monitoring instruction in classrooms),
- good teaching builds on students strengths and respects individual differences,
- good teaching involves modeling what students should learn
- general references to teachers' use of effective teaching and learning practices (this includes discussions of teachers' use of best practices)

Other dimensions might include but are not limited to:

- cognitively or developmentally appropriate or challenging curriculum for students
- applied learning theory
- individualized instruction

- reciprocal teaching
- inquiry teaching or direct instruction

1. A Little

Mere mention of *one or two aspects* of effective teaching and/or learning with no development of the aspect(s). NOTE: mentioning the same thing 10 times with no development is still a mere mention.

2. Some.

Mentions at least *three or more different aspects* of effective teaching and learning but does not develop any of the aspects.

3. Sufficient

Mentions at least *one* aspect of effective teaching and learning and develops at least *one aspect*; that is, the response goes beyond mention of an aspect to develop it suggesting a deeper understanding. (For example, the respondent might mention effective instructional strategies in reading and say teachers need to use "writing workshop" or "balanced literacy." Or, the respondent might mention evidence based teaching or assessment and go on to note trying to figure out the strategies that teachers use who have high performing students).

Specific example of single aspect (individualized instruction) that is developed:

> *"Students must have pre assessment in the critical areas of reading such as vocabulary, phonics, fluency, comprehension, etc. Teachers must know the basic reading levels of their students. Instruction must be tailored to meet these specific needs."*

4. Quite a Bit

Mentions at least *two* aspects of effective teaching and learning and develops *two or more*; that is, the response goes beyond mentioning the aspects to developing them with more discussion that suggests a deeper understanding of the aspects.

5. A Great Deal

Mentions at least *two* aspects of effective teaching and learning and develops *two or more AND makes connections* between at *least two* of the aspects mentioned; that is, the response goes beyond mentioning and developing two or more aspects of effective teaching and learning to making a link or connection between at least two aspects. For example, the respondent might mention and develop how student motivation is critical and then link it to how student effort produces achievement rather than IQ alone. A second example could be that a principal develops 1) how to determine if teachers are using best practices in their teaching, and 2) the importance of using individualized instruction, and she/he then connects them by discussing how individualized instruction should be included as a part of best practices.

References

Birckmayer, J. D., & Weiss, C. H. (2000). Theory-based evaluation in practice: what do we learn? *Evaluation Review, 24*(4), 407–431.

Bloom, H. S. (Ed.). (2005). *Learning more from social experiments: Evolving analytic approaches*. New York: Russell Sage Foundation.

Borman, G. D., Slavin, R. E., Cheun, A., Chamberlain, A. M., Madden, N. A., & Chambergs, B. (2005). Success for all: first-year results from the national randomized field trial. *Educational Evaluation and Policy Analysis, 27*(1), 1–22.

Boruch, R. (1997). *Randomized experiments for planning and evaluation: A practical guide*. Thousand Oaks: Sage.

Boruch, R. (2002). The virtues of randomness. *Education Next, 2*(3), 36–41.

Brewer, J., & Hunter, A. (1989). *Multimethod research: A synthesis of styles*. Thousand Oaks: Sage.

Bryman, A. (2006). Integrating quantitative and qualitative research: how is it done? *Qualitative Research, 6*(1), 97–113.

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

Caracelli, V. J., & Greene, J. C. (1997). Crafting mixed-method evaluation designs. In J. C. Greene & V. J. Caracelli (Eds.), *Advances in mixed methods evaluation: The challenges and benefits of integrating diverse paradigms*. San Francisco: Jossey-Bass.

Chatterji, M. (2005). Evidence on "what works": an argument for extended-term mixed-method (ETMM) evaluation designs. *Educational Researcher, 34*(5), 14–24.

Chen, H. T. (1990). *Theory-driven evaluations*. Newbury Park: Sage.

Chen, H. T. (1997a). Applying mixed methods under the framwork of theory-driven evaluations. *New Directions for Program Evaluation, 1997*(74), 61–72.

Chen, H. T. (1997b). Normative evaluation of an anti-drug abuse program. *Evaluation and Program Planning, 20*(2), 195–204.

Chen, H. T. (2005). *Practical program evaluation: Assessing and improving planning, implementation and effectiveness*. Thousand Oaks: Sage.

Chen, H. T., & Rossi, P. H. (1980). The multi-goal, theory-driven approach to evaluation: a model linking basic and applied social science. *Social Forces, 59*(1), 106–122.

Chen, H. T., & Rossi, P. H. (1983). Evaluating with sense. *Evaluation Review, 7*(3), 283–302.

Chi, M. T. H. (1997). Quantifying qualitative analyses of verbal data: a practical guide. *The Journal of the Learning Sciences, 6*(3), 271–315.

Cook, T. D. (2002). Randomized experiments in educational policy research: a critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis, 24*(3), 175–199.

Cook, T. D., & Reichardt, C. S. (Eds.). (1979). *Qualitative and quantitative methods in evaluation research*. Thousand Oaks: Sage.

Cook, T. D., & Wong, V. (2006). *The IES agenda to institutionalize randomized clinical trials in educational research: Description and commentary*. Paper presented at the Institute for Policy Research Spring 2006 Colloquium.

Cook, T. D., Murphy, R. F., & Hunt, H. D. (2000). Comer's school development program in Chicago: a theory-based evaluation. *American Educational Research Journal, 37*(2), 535–597.

Creswell, J. W. (2002). *Educational research: Planning conducting, and evaluation quantitative and qualitative research*. Upper Saddle River: Merrill Prentice Hall.

Creswell, J. W., Shope, R., Plano Clark, V. L., & Green, D. O. (2006). How interpretive qualitative research extends mixed methods research. *Research in the Schools, 13*(1), 1–11.

Cronbach, L. J. (1988). Five perspecitves on the validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity*. Hillsdale: Erlbaum.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Intelligence: Measurement, theory, and public policy*. Urbana: University of Illinois Press.

Datta, L. (1994). Paradigm wars: a basis for peaceful coexistence and beyond. *New Directions for Program Evaluation, 61*, 53–70.

Denzin, N. K. (1978). *The research act: A theoretical introduction to sociological methods* (2nd ed.). New York: McGraw-Hill.

Denzin, N. K. (1989). *The research act: A theoretical introduction to sociological methods* (3rd ed.). Englewood Cliffs: Prentice-Hall.

Denzin, N. K., & Lincoln, Y. S. (Eds.). (2005). *The SAGE handbook of qualitative research* (3rd ed.). Thousand Oaks: Sage.

Eisenhart, M., & Towne, L. (2003). Contestation and change in national policy on "scientifically based" education research. *Educational Researcher, 32*(7), 31–38.

Flemming, K., Adamson, J., & Atkin, K. (2008). Improving the effectiveness of interventions in palliative care: the potential role of qualitative research in enhancing evidence from randomized controlled trials. *Palliative Medicine, 22*(2), 123–131.

Goldring, E., Huff, J., Spillane, J. P., & Barnes, C. A. (2009). Measuring the learning-centered leadership expertise of school principals. *Leadership and Policy in Schools, 8*(2), 197–228.

Gottlieb, N. H., Lovato, C. Y., Weinstein, R., Green, L. W., & Eriksen, M. P. (1992). The implementation of a restrictive worksite smoking policy in a large decentralized organization. *Health Education & Behavior, 19*(1), 77–100.

Greene, J. C. (2006). Toward a methodology of mixed methods social inquiry. *Research in the Schools, 13* (1), 93–98.

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.

Guba, E. G. (1990). The alternative paradigm dialog. In E. G. Guba (Ed.), *The paradigm dialog*. Newbury Park: Sage.

Hall, B., & Howard, K. (2008). A synergistic approach: conducting mixed methods research with typological and systemic design considerations. *Journal of Mixed Methods Research, 2*(3), 248–269.

Howe, K. R. (1988). Against the quantitative-qualitative incompatibility thesis or dogmas die hard. *Educational Researcher, 17*(8), 10–16.

Howe, K. R. (2004). A critique of experimentalism. *Qualitative Inquiry, 10*(1), 42–61.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: a research paradigm whose time has come. *Educational Researcher, 33*(7), 14–26.

Klein, G. A., Calderwood, R., & MacGregor, D. (1989). Critical decision method for eliciting knowledge. *IEEE Transactions on Systems, Man, and Cybernetics, 19*(3), 462–472.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: confirmation from meta-analysis. *American Psychologist, 48*(12), 1181–1209.

Lynch, S., Szesze, M., Pyke, C., & Kuipers, J. (2007). Scaling up highly rated middle science curriculum units for diverse student populations: Features that affect collaborative research and vice versa. In B. Schneider & S.-K. McDonald (Eds.), *Scale-up in education: Issues in practice* (Vol. 2). Plymouth: Rowman & Litlefield.

Mactavish, J. B., & Schleien, S. J. (2004). Re-injecting spontaneity and balance in family life: parents' perspectives on recreation in families that include children with developmental disability. *Journal of Intellectual Disability Research, 48*(2), 123–141.

Mathison, S. (1988). Why triangulate? *Educational Researcher, 17*(2), 13–17.

Maxwell, J. A., & Loomis, D. M. (2003). Mixed methods design: An alternative approach. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research*. Thousand Oaks: Sage.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity*. Hillsdale: Erlbaum.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*. Thousands Oaks: Sage.

Morgan, D. L. (1998). Practical strategies for combining qualitative and quantitative methods: applications to health research. *Qualitative Health Research, 8*, 362–376.

Morse, J. M. (1991). Approaches to qualitative-quantitative methodological triangulation. *Nursing Research, 40*(2), 120–123.

Morse, J. M. (2003). Principles of mixed methods and multimethod research design. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research*. Thousand Oaks: Sage.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: implications for performance assessment. *Review of Educational Research, 62*(3), 229–258.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher, 23*(2), 5–12.

Nastasi, B. K., & Schensul, S. L. (2005). Contributions of qualitative research to the validity of intervention research. *Journal of School Psychology, 43*(3), 177–195.

National Policy Board for Educational Administration (2002). Standards for advanced programs in educational leadership for principals, superintendents, curriculum directors and supervisors.

Niglas, K. (2004). *The combined use of qualitative and quantitative methods in educational research*. Tallinn: Tallinn Pedagogical University.

Petrosino, A. (2000). Answering the "why" question in evaluation: the causal-model approach. *Canadian Journal of Program Evaluation, 15*(1), 1–24.

Ragin, C. C. (2000). *Fuzzy-set social science*. Chicago: University of Chicago Press.

Reichardt, C. S., & Rallis, S. F. (Eds.). (1994). *The qualitative-quantitative debate: New perspectives*. San Francisco: Jossey-Bass.

Riggin, L. J. C. (1997). Advances in mixed-method evaluation: a synthesis and comment. *New Directions for Program Evaluation, 74*, 87–94.

Rossi, P. H., Lipsey, M. W., & Freeman, D. J. (2004). *Evaluation: A systematic approach* (7th ed.). Thousand Oaks: Sage.

Rossman, G. B., & Wilson, B. L. (1993). Numbers and words revisited: being "shamelessly eclectic". *Quality and Quantity, 28*(3), 315–327.

Sandelowski, M. (1996). Focus on qualitative methods: using qualitative methods in intervention studies. *Research in Nursing and Health, 19*(4), 359–364.

Sandelowski, M. (2003). Tables or tableaux? The challenges of writing and reading mixed methods studies. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social dn behavioral research*. Thousand Oaks: Sage.

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theories of practice*. Newbury Park: Sage.

Shavelson, R. J., & Towne, L. (Eds.). (2002). *Scientific research in education*. Washington, DC: National Academies Press.

Spillane, J. P., Camburn, E. M., & Pareja, A. S. (2007). Taking a distributed perspective to the school principal's workday. *Leadership and Policy in Schools, 6*(1), 103–125.

Strauss, A. L., & Corbin, J. M. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Thousand Oaks: Sage.

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

Tashakkori, A., & Teddlie, C. (Eds.). (2003). *Handbook of mixed methods in social and behavioral research*. Thousand Oaks: Sage.

Weiss, C. H. (1997). How can theory-based evaluation make greater headway? *Evaluation Review, 21*(4), 501–524.

Yin, R. K. (2006). Mixed methods research: are the methods genuinely integrated or merely parallel? *Research in the Schools, 13*(1), 41–47.