Promoting Learning in Organizations Through Embedded Case Studies

Stephen Newton

Program Evaluation and Research Branch
Los Angeles Unified School District

Annual Meeting of the American Evaluation Association
November 2003 Reno, Nevada
Abstract

The use of embedded case studies can complement more general data collection efforts in evaluation. In the context of results from a more general data collection effort, such theories can be a useful complement to enhance the depth of understanding of key issues of interest to stakeholders. Three examples of evaluations conducted in a large urban school district illustrate the use of this method, suggest obstacles, and demonstrate the usefulness of this approach. The method used in these three evaluations utilized embedded case studies to develop explanatory theories about key aspects of programs. First, a study of a coaching program for teachers used teachers as cases and found that certain skills used by the coach differentiated successful from unsuccessful outcomes. Second, a study of a math reform initiative analyzed lessons from different levels of schooling as cases, and showed that teachers did not address issues of conceptual understanding of mathematics. Third, a study of career academies, high school small learning communities, found that the personal relationships were a key factor in supporting students to succeed. Case studies can be used to develop and test plausible explanatory hypotheses regarding key features of a program. Within the context of promoting organizational learning (Preskill and Torres, 1999) such inferences can provide the starting point for addressing key issues related to a program’s functioning.
Promoting Learning in Organizations Through Embedded Case Studies

This article presents a method for using embedded case studies as a way to complement more general data collection efforts. Three examples of evaluations conducted in a large urban school district illustrate the use of this method, suggest obstacles, and demonstrate the usefulness of this approach. The method used in these three evaluations utilized embedded case studies to develop explanatory theories about key aspects of programs. In the context of results from a more general data collection effort, such theories can be a useful complement to enhance the depth of understanding of key issues of interest to stakeholders. By providing this type of information, this approach serves to increase the informativeness of evaluations, and therefore their usefulness.

The term “embedded case study” typically refers to a case that is embedded within a larger case study (Scholz & Tietje, 2003; Yin, 2003). In this paper, embedding of case studies refers either to locating multiple case studies within a larger case study or locating them within any broad depiction of a program, which could be the result of qualitative and/or quantitative methods. The important issue is that the case studies are located within a broader perspective on the program, and in this sense are “embedded.”

Before discussing the conceptual and practical issues related to using case studies in this way, three examples are presented in which evaluators embedded case studies within larger evaluations in a large urban school district, two of which were done on
relatively small projects by this evaluator and one in a larger evaluation project done by a colleague (Ai, November 2002).

Examples

*Peer Assistance and Review Program*

The first evaluation studied a Peer Assistance and Review (PAR) program (Newton, August 2002). Since this was a fairly new program in the first year of evaluation, stakeholder concerns reflected a desire for broad feedback to refine the program. The PAR program provides coaches to either veteran teachers that had been rated by their administrator as below standard in the previous year, or new teachers at poor-performing schools. The evaluation design involved using multiple methods to study the program as a whole, and especially a linked sample of teachers, coaches, and administrators, observations of trainings, and document review. The first year’s report first summarized descriptive information about the program’s functioning, and the overall outcomes for participants. The second section of the report treated each teacher as a case, and explored the data to determine themes that distinguished teachers who showed substantial improvement from those who did not. The report developed an explanatory theory that successful coaching relationships required an acceptance and understanding of normative teaching standards, careful observation by the coach, and incremental feedback. Because there was a concern by program staff for confidentiality due to the sensitive personnel issues involved, the cases could not be reported in enough depth so that teachers could be identified. Therefore, the theory was illustrated with information from the case studies, but cases were not presented in detail in the report. As a result of this evaluation, program staff made use of the findings from this report both in planning
the overall program and in trainings for coaches. Directors of the program implemented changes that incorporated teaching standards throughout the program, which was consistent with the theory about effective coaching presented in the first year’s report.

A follow-up evaluation the next year followed a similar design and found the same patterns to characterize successful relationships, but refined the concepts based on new information (Newton, September 2003). For example, in the first year when coaches emphasized extrinsic reasons for teachers to change their teaching (typically to please the principal), they were more likely to make only shallow or inconsistent change in their teaching. In the second year, some teachers improved dramatically even when their coaches placed some emphasis on pleasing the principal, as long as their coaches placed primary emphasis on teaching standards. Another finding in the second year was that fewer teachers were reported to make shallow or inconsistent change than in the first year’s evaluation. This result was consistent with the theory that motivating teachers to teach according to teaching standards was a key feature of effective coaching, since the program had begun to emphasize teaching standards throughout, and this was the predicted result from such a change.

Mathematics Reform Initiative

Xiaoxia Ai studied a math reform program in the school district that included new curriculum, instructional coaching, and training throughout all levels of school (Ai, November 2002). This evaluation used a logic model to summarize the presumed impact of reform in the district’s mathematics program. Broad data regarding test results, textbooks, training, and coaching were collected to depict the overall picture of reform in the district. The evaluation also included observations of 160 classrooms involving
almost 1,000 lessons, with attention paid to carefully record the instructional dialogue between students and teachers. This analysis summarized the discourse in classrooms addressing the concept of “place value” in grades 2, 4, 8, and 10. These classrooms reflected practice that was observed across the entire sample, but the cases studies were selected because they had commonality of subject matter that was treated at different levels of schooling. The analysis documented how teachers followed a discourse norm that focused on procedural knowledge and failed to create opportunities to learn mathematical concepts. Students were unable to construct mathematically valid mathematical arguments even in honors classes in the upper grades. It also showed that students at these upper grades did not have a solid grasp of the concept of place value that could have been addressed beginning in early grades, but was not. In the report, these findings are juxtaposed with the overall findings regarding the program. This in-depth depiction of teaching practice and student understanding complemented the more general information found about the program, and was not synthesized explicitly. Rather, the findings regarding classroom dialogue were presented in juxtaposition to the other findings regarding the program as a whole. The report section analyzing classroom dialogue was later found to be useful for training teachers and coaches.

*High School Career Academies*

A third evaluation studied high school career academies with the theme of teaching as a career (October, 2003). Program staff requesting the evaluation expressed interest in obtaining information that could be used to help people develop new academies at other schools in the district. Very little was known about the actual functioning of these academies at the time the evaluation was requested, and as was
found that only 13 of the 15 academies to be studied were actually in existence. Data collection was designed in two phases, one to get an overall picture of the functioning of the academies, and the second involved case studies of successful academies. The initial data collection indicated that most often the higher implemented academies were at lower-performing schools, so case study schools were selected to reflect different levels of overall school performance, to see how this might contribute to the challenge of developing a successful academy. Since this report is currently under review, the findings cannot be fully presented here, but the report synthesized the overall descriptions of the programs with the in-depth study of specific schools that included observations of various academy activities and interviews with a sample of all types of participants, including recent graduates. These case studies helped to ground the more general descriptions in a much more specific, nuanced way, utilizing the small learning community literature to develop an understanding of what makes academies function effectively. This report is being used as a component for a training of academy coordinators.

Overview

In the evaluations described above, various approaches were used to develop broad descriptions of the programs, including use of test scores, interview data, and survey data. General conclusions were derived regarding the nature of each program and its effects on participating students and teachers, using such methods. It would be an oversimplification to identify quantitative methods with breadth and qualitative with depth, because methods of each type can address the converse issue as well. But as a
gross generalization, quantitative methods are more easily suited for depicting information regarding a program as a whole.

The embedded case study method described herein assumes a broader study of the program as a whole within which to embed the case studies. This broader method should address the overall nature of the program, and ideally evaluate the effectiveness of the program in achieving worthwhile goals. Within the context of using this method, multiple case studies can provide a particularly helpful way to address key program issues, by capitalizing on the holistic approach to studying a phenomenon, and the use of replication logic (Yin, 2003) to generate plausible explanatory theories. Case studies allow for a range of information that applies to a phenomenon to be taken into consideration all at once. Since a case is by definition a single entity, findings cannot be generalized to statistically represent other entities. This, however, should not be treated as a shortcoming, because the logic of embedded case studies and the epistemological underpinning upon which it is grounded differ from that of statistical methods.

The logic of embedded case study analysis follows Yin’s (2003) rationale that inferences from multiple case studies reflect a “replication logic.” Such a logic generalizes to theory rather than to a population. That is, analysis of multiple cases allows for the generation of a theory consistent with the case data. Such a theory based on the embedded case study analysis cannot be automatically generalized to a population based on this approach. However, depending on the strength of the theory (importance, robustness of the results, characteristics of the cases in comparison with characteristics of those to which theory will be generalized), it can prove useful for helping guide program decisions as decision makers weigh these factors.
The underlying understanding of knowledge and the nature of scientific inquiry in this method is rooted in Polanyi’s philosophy of science (1958). Polanyi argues that all knowledge is “personal knowledge” and must inevitably be grounded in tacit knowledge. The concept of tacit knowledge has been influential in the conceptualization of qualitative research (Lincoln & Guba, 1985). Fewer thinkers have placed as much emphasis on Polanyi’s view that all knowledge, even scientific knowledge, must therefore be personal, and requires scientists to have personal responsibility for what they think and believe. Furthermore, it should be noted that these concepts were developed to describe the ways that physical scientists (Polanyi was a chemist) developed their knowledge.

Consistent with Polanyi’s theory, the embedded case study method assumes that stakeholders have substantial tacit knowledge of their programs, and that the types of information provided by evaluations will be interpreted in the light of their explicit and tacit knowledge. The in-depth investigations provided by case studies allow the researcher to develop depth of tacit knowledge about the program they are researching, and can therefore help to shape the analysis to take into account such tacit knowledge.

Embedded Case Study Method

Design

The embedded case study method requires several steps. First, the methods for selecting cases should be considered, but not necessarily finalized, before data collection begins. The value of the information that can be derived from case studies depends heavily on the appropriateness of the case selection for the purposes of the study. Typically, case study selection is made by purposive sampling rather than probability
sampling because case studies are not intended to generate statistical generalizations. The purpose of multiple case studies is to use selected cases in combination to generate and test key explanatory theories. Therefore, they should be selected so as to provide the strongest ways to generate such theories. Cases that are too divergent in many parameters might be difficult to compare with each other in any meaningful way. Similarly, cases that are too similar might not provide the opportunity to rigorously test the hypotheses generated.

In the examples of the three evaluations described earlier, two different methods to select cases were used. In the study on the PAR program and the math reform, a large number of cases had data collected, and a subsample was selected to address the key issue of concern. In the PAR study, teachers who made dramatic change were contrasted with teachers who made shallow change and a profile of different coaching practices was developed. In the math study, cases were selected that dealt with similar subject matter across different grade levels. In both cases, the cases examined reflected broader patterns within the data, and this was discernible because of the large numbers of cases. In the third evaluation, case selection was planned to be the second phase of data collection, because the case studies of whole programs required substantial resources and therefore could not be done at every school. Since the focus was on successful practices, cases were selected that appeared to be highly implemented. The initial data collection suggested that programs in lower performing schools tended to be larger and more successful than ones in higher performing schools, so case studies were selected from successful program across different levels of school performance, with the idea that they would be similar, but perhaps would help explain the reason for the pattern regarding the
success or failure in different contexts. One challenge in selecting cases based on initial data collection is related to timing. Selecting cases based on data from a general data collection effort requires that the initial data collection be complete and analyzed before selecting and initiating case studies. Because data collection and analysis must therefore be conducted sequentially, careful attention should be paid in planning in order to complete everything in time.

Analysis

Data analysis of the larger data collection and case studies is completed independently according to the methods appropriate to each. In each of these cases, the larger data collection was guided by a theory-driven approach that used a logic model to depict program components and their interrelationships, and data was collected to match this model. The larger analysis thereby provided an overall account of the program elements and main outcomes that were seen to result from the program whereas embedded case studies are focused on developing a more circumscribed theory about key features of a program.

The relationship between case study findings and the larger analysis can be addressed in several ways. Since the modes of analysis differ, with the general data collection used to depict the empirical features of the program as a whole and case studies used to develop key explanatory theories related to only a part of a program, the types of conclusions reached by each method should not be combined without considerable care.

Case study analysis provides the opportunity to explore the phenomenon of interest in a holistic manner. The richer and more complete the data, the stronger the case
study. Case studies can function to suggest plausible causal/explanatory hypotheses about a program, and can also serve as points of comparison to see if such hypotheses are consistent with the data. Replication logic refers to using each case as a test of a hypothesis, much as scientific studies can be done to replicate other studies, to see if they find similar findings. Yin (1993) requires that hypotheses be specified in advance of data collection, and then tested in this manner. However, while this approach has the advantage of clarity in distinguishing data that helped generate a hypothesis from data used to test a hypothesis, it is not clear that such rigor is absolutely necessary.

The concern for predetermining hypotheses prior to examining data addresses the problem that a hypothesis can be generated that merely correspond with the data, and therefore replication does not provide a genuine test. However, as a practical matter, there are many other factors that constrain the development of plausible hypotheses. Hypotheses should make some account of how the program leads to outcomes, or at least a substantial part of this issue. For most programs, the set of outcomes is quite limited and so the explanatory theory must connect program features with this small set of positive indicators. For example, in educational settings, student learning is a typical outcome, so the hypothesis would need to address either student learning or a factor strongly related with it. Second, programs also have few elements that are likely to lead to such outcomes. For example, the PAR program had a goal of improving teachers instruction to an acceptable level according to the California Standards for the Teaching Profession. Case study analysis that seeks to explain how coaches help teachers to do this is constrained by making an account of what distinguishes such outcomes for teachers.
Furthermore, explanations are also constrained by the fact that they must be plausible, and this criterion should be consistent both with research literature and the experiential knowledge of practitioners. For example, if it were found that only coaches with birthdays in the Summer were effective, this would probably not be persuasive as a plausible account of how to select effective coaches. Given such constraints, and the opportunity to use multiple cases to take into account to develop detailed account of key aspects of a program, it seemed legitimate to develop plausible explanatory hypotheses (Stake, 1993) by using case studies. The strength of the explanation should be weighed by considering a host of issues that could affect the legitimacy of the account, and when reasonable alternative explanations are available, they should be mentioned. An exhaustive description of this reasoning process is not possible in this space, but thinkers such as Popper and Campbell provide some help. The potential value of presenting explanations typically outweighs the danger of overstepping the risk of being wrong in most situations. To the extent that an erroneous inference would have a negative impact, say for example, ending a program altogether, then the inference should be subjected to more rigorous tests and presented more tentatively.

Reporting

The findings from case studies in these three evaluations were presented separately from the findings from other data collection in each of the reports described here. The main relationship between the findings from different aspects of these reports is that they are complementary, using different analyses to reach different types of conclusions. The general approaches developed an empirical description of the program as a whole, whereas the case studies focused on holistic in depicting the phenomena as
clearly as possible as they actually existed, and also extracted theoretical explanations that helped to account for the overall patterns. The findings from the case studies were both complementary to the other findings, and as more narrowly focused methods, contextualized within the larger empirical picture presented by the more general data collection method. For example, in the study of career academies, it was found that academies varied quite a bit in their level of implementation. Students at high implementation academies outperformed their peers at the same schools quite substantially, and this difference grew as compared to their academic performance prior to high school. The case studies showed that at high implementing academies, students and teachers felt like they knew each other personally, and triangulation of data showed that students were encouraged in numerous ways to succeed in school and go to college. Data from the general data collection showed that lower implementing academies did not incorporate many of the ways of building relationships between participants, and also did not have the same positive outcomes. This suggested that the communal features of the academy played an important part in their success, and helped to explain the difference between outcomes for high and low-implementing academies.

It is also an option to not closely integrate the findings from the different methods, but instead to juxtapose them. This leaves it to the reader to develop interpretations that can account for the different patterns of results. For example, in the mathematics reform study, the overall pattern of scores showed that students did not perform well on standards tests, especially in upper grades. The case studies of classrooms, as described earlier, showed how students were not given the opportunity to learn a conceptual understanding of mathematical concepts. The case studies powerfully
suggested the importance of conceptual understanding, but the inference was left to the reader since the link was not established in a powerful empirical manner (e.g., correlation analysis or regression), though conceptually a compelling case could be made about its importance.

Regardless, care must be taken to present findings from the different methods in a way that remains true to their strengths and weaknesses. The information that can be obtained through case studies (holistic, specific, and suggestive of powerful explanations of phenomena) can be highly complementary to general findings about a program. Juxtaposition of these results can easily be suggestive of important ways to understand a program. More extensive integration of findings from the different methods requires additional care, but can also be worthwhile.

Summary

Case studies can thus result in presenting plausible explanatory hypotheses regarding key features of a program. Within the context of promoting organizational learning (Preskill and Torres, 1999), such inferences can provide the starting point for addressing key issues related to a program’s functioning. For example, in the study of the PAR program, the key importance of well-developed coaching skills provides a deeper perspective on what it might take to have an optimal program. Similarly, in the evaluation of a math reform program, the difficulty of teaching for conceptual understanding and its apparent effects on the ongoing development of students raises important questions about the depth of reform needed to teach students to perform at a high level of mathematics. Finally, in the study of career academies, the importance of small learning community features suggests the importance of personalizing education.
In none of these evaluations were these theories fully established empirically. The strength of the findings depends on the numbers of cases analyzed, the consistency between the theory and other relevant theories, and the degree to which alternative explanations were ruled out by the data. Further consideration should be made in determining the extent to which such inferences are safe in high stakes evaluations. There is little doubt, though, that such case studies can be very informative and therefore useful to stakeholders.
References


