

 **VIGNETTE: Documenting Treatment With Implementation Scales**

Purpose: When you evaluate the implementation of your magnet program, it's sometimes important to measure the degree to which various schools are engaging in treatment activities. Doing so can help reveal otherwise hidden differences in breadth and depth of program implementation between program and comparison schools. Such scales can be especially useful when the two schools have several program elements in common. Reflecting on this vignette may suggest how "implementation scales" would be useful if you face a similar situation of overlap between magnet and comparison school treatment.

Note: Similar scales can be used to determine whether participants are partially or fully implementing a program, and can account for differences in implementation that often occur between year one and year two. Scales can also be used in rigorous evaluations to confirm a unique treatment of magnet schools relative to comparison schools.

Source: Interview with Naida Tushnet, evaluator with WestEd, December 2, 2008.

Questions for Reflection

1. Naida Tushnet mentions that her evaluation team developed scales for 11 different elements of program implementation, including parent involvement and use of research-based reforms. For which of your program components might you develop scales? Why?
2. The scales that Naida Tushnet and her team developed utilized surveys as one set of measures. What other measures could be used as part of your scales?
3. Once measures have been selected, how will you determine what constitutes "emerging" versus "full" implementation?
4. Do you expect differences between year one and year two of program implementation? How can the scale highlight these differences?
5. How will you ensure that those who fill out the scale are well calibrated on what constitutes a 1, 2, or 3?

Documenting Treatment With Implementation Scales

Background: Naida Tushnet, an evaluator with WestEd, and her colleagues conducted the national evaluation of the Comprehensive School Reform (CSR) program (2002–present). The team was charged with documenting the relationship between program implementation and outcomes. Because the patterns of implementation between CSR and comparison schools were inconsistent, they designed the scale to measure the degree of implementation so they could determine in what ways outcomes were associated with the program (CSR).

As much as we try to find comparison sites that don't share elements of a program site, it's sometimes very difficult to do so and still match on other things such as demographics and prior achievement. For example, in the program we were evaluating we were looking at a specific set of 11 reform activities. If you looked at our comparison schools, because they were also Title I schools, they had a number of activities going on that were the same as the program schools. One activity was parent involvement. Well, all Title I schools have parent involvement programs. Another component was use of scientifically based curriculum and instruction. But under No Child Left Behind, everybody is supposed to use scientifically based curriculum and instruction. And, in fact, a number of the comparison schools were using similar reading and math programs as those that the treatment schools were using. So, then you have the question: With whom do we compare these program schools if their treatment is not really unique?

So, we constructed a scale to measure **the degree of program implementation**. We hypothesized that we would see more of the 11 reform elements in the program schools than we would in the comparison schools, and that fuller implementation would lead to better outcomes. By using a scale showing degree rather than just a yes-no option, you could more closely examine and judge whether the treatment made a difference.

Figure 1: Component 3 and 7 of Implementation Scales

Component	Measure			Score
3. Professional Development				
3.1 Strong content focus:	Yes		No	1
3.2 Use of non-traditional teaching strategies:	Yes		No	1
3.3 Range of PD hours required or taken by average teacher per year:	54+	36-53	Aug-35	54+ = 3 36 - 53 = 2 8 - 35=1
3.4 Aligned with the student needs/data-driven instruction	Yes		No	1
3.5 Evidence of collective participation in groups of teachers within the same school:	Yes		No	1
3.6 Evidence of some PD taking place in the teacher's classroom:	Yes		No	1
3.7 Explicit guidance to align PD with standards, curriculum, or assessment tools:	Yes		No	1
Total possible score for Component 3				9

Component	Measure		Score
7. Parent and Community Involvement			
7.1 Emergence of new forms of parent involvement during CSR years:	Yes	No	
7.1.1 Special parent events:	Yes	No	3 – 4 = 1 0 – 2 = 0
7.1.2 Programs or opportunities for parents in instructional roles:	Yes	No	
7.1.3 Parent Advisory or other committees	Yes	No	
7.2 Level of Parental involvement (high, medium, or low, defined as follows):			
High: You've observed parents in the school and interviewees voice strong or satisfactory level of parental involvement in school activities		High	High = 2
Medium: Explicit guidance to align PD with standards, curriculum, or assessment tools		Medium	Medium = 1
Low: no evidence of parental involvement beyond a handful of parents and interviewees voice low levels of participation		Low	Low = 0
7.3 Evidence of at least one community organization and one school/community event or program:	Yes	No	1
Total possible score for Component 7			4

We constructed the scale with each of the 11 reform components in it and developed measures for each of those components. For example, we looked for evidence of parents' involvement in children's learning through site visit observations of whether parents volunteered in classrooms, sign-in sheets for parent-teacher events, the presence of a parent resource center, and the existence of a family literacy program. These observations were noted on the scale and rated. The summary rating (high, medium, low) and the other items on the scale were used to rate implementation and relate it to outcomes in both program and comparison treatments. For the element "use of scientifically based programs," one descriptor might be related to how the decision was made to select a particular curriculum or instructional approach. Then a question in a survey would ask which reasons were most important in how you adopted your program, and to what extent were you involved in the adoption decision?

Developing a scale like this is certainly not something we could have done without the right people on board and sufficient time and funding. Another issue that comes up when you use a scale is what do you do when the principal responds to a survey question in one way and the teachers' response to that question is different? Or, what if observations don't align with reports from staff? For example, an observer may have actually counted the attendance at the parent meetings and reported five parents at one meeting and 10 at another. This might not match the self-reported data, such as a teacher's survey response that 50% of the parents showed up at meetings.

Once we developed the scale, we could determine whether schools that scored higher on the implementation scale produced better outcomes than those that did not. And because the treatment schools had a deeper or broader implementation of the 11 components, we could say something about how the program actually makes a difference.