**Title**

Systematizing the Measurement and Reporting of Intervention Delivery in Education Research

**Authors**

Catherine Darrow, Ph.D.
Barbara Goodson, Ph.D.
Beth Boulay, Ed.D.

Abt Associates, Inc.
Cambridge, MA
**Background / Context**

Education researchers and evaluators have established the need to measure levels of implementation as a means to confirm delivery (Gearing et al., 2011) and interpret program effects (Cordray, 2009). However, the field remains uncertain of the methods by which to effectively and systematically collect and pool implementation data in a meaningful way that can be used to contextualize impact estimates. Without a clear system of measurement, evaluators are unable to quantify levels of fidelity in a manner that can be used to both monitor program implementation more precisely and to provide evidence that adequate inputs and resources were provided to the deliverers of the intervention to be able to implement the intervention with fidelity.

There are a number of reasons for measuring the fidelity of any program which include but are not limited to monitoring and ensuring delivery of a program (Schoenwald, et al., 2011) and utilizing fidelity data to validate claims of program effectiveness (Dusenbury, Brannigan, Falco, & Hansen, 2003). Measurement of program delivery is one means to confirm that the critical components of an intervention, originally designed by the developer and financially supported by a funding agency, are in fact being delivered as planned to the intended target population. Researchers may use data collected via comprehensive fidelity measures to report the degree to which the program is adopted by practitioners and provided to recipients. Resulting analyses can be provide to funding agencies as evidence of program implementation and can also be used to allow implementers an opportunity for quality improvement (Bond, Drake, Rapp, McHugo, & Xie, 2009).

The question of program effectiveness and related concerns of program implementation have come to the forefront of educational research. In recent years, funding agencies, like the United States Department of Education (ED) have required grant recipients to measure the fidelity of implementation of interventions as part of rigorous impact studies. Measuring fidelity requires clear specification of the program model being tested, including identifying critical components of the model, their relationship to each other, and their relationship to intermediate and long-term outcomes (Bond, Evans, Salyers, Williams, & Kim, 2000; Century, Rudnick, and Freeman, 2010; Hall and Hord, 1987; Mowbray, Holter, Teague, & Bybee, 2003).

Discussion within the field has pulled from Dane and Schneider’s seminal review (1998) of the ways in which evaluation research reveals varying levels of implementation and has further developed a number of approaches to measuring implementation. It can be argued that consensus has been reached over the need for program developers, researchers, and evaluators to clearly define the program model (Century, Rudnick, & Freeman, 2010; Nelson, Cordray, Hulme, Darrow, & Sommer, 2012), to indicate the critical elements that are theoretically tied to impacts of interest (Mowbray, Holter, Teague, & Bybee, 2003; O’Donnell, 2008), to develop original or use pre-existing measures to assess levels of fidelity (Century, Cassata, Rudnick, & Freeman, 2012; Hume et al., 2011), and for using fidelity data to interpret program impacts (Durlak & Dupre, 2008; Hulme & Cordray, 2009).

Ongoing developments in implementation science are impressive, particularly in the ways education researchers have devoted attention to measuring fidelity to specific interventions. A wealth of approaches have emerged including methods in measuring fidelity to targeted
interventions (see Hume et al. (2011) and Pence, Justice, & Wiggins (2008) for examples). With this influx of examples of fidelity measurement comes a pool of approaches that are tied to specific interventions which assess the degree to which unique elements of an intervention are implemented. Diverse approaches in defining and measuring fidelity abound (O’Donnell, 2008). Each system of measurement is seemingly developed and presented in isolation with few connections across educational interventions. Yet, a universal approach in collecting and pooling implementation data, often a large amount from varying sources, into quantifiable, analyzable values is underdeveloped. No systematic method of measurement, analysis, and reporting that enables fidelity results to be compared across different interventions and across evaluations exists. This presentation adds to the field by offering insight into a process of systematically combining fidelity data to represent the implementation of any complex intervention.

Purpose / Objective / Research Question / Focus of Study

The proposed presentation and associated manuscript contributes to a growing body of knowledge related to implementation science, by first, clearly defining various approaches in measuring fidelity and then describing the current state of fidelity measurement in educational evaluation research. It will then present an approach to developing of fidelity of implementation measurement system and combining data through a multi-step process. This process urges evaluators to consult the intervention logic model, operationalize the identified key components, create thresholds to determine extent of implementation, and synthesize results across indicators, components, and units of analysis (e.g., classrooms, schools) to a sample-level calculation. Several considerations emerge throughout this process related to decisions around intervention implementation expectations and objectives, accurate and meaningful scoring systems, and theoretical decisions about implementation dosage to answer how much of the intervention is enough to expect changes in the intended outcomes.

Setting / Population / Participants / Subjects

To most effectively illustrate the application of this approach to measurement, analysis, and reporting of implementation, data from the Massachusetts ELT evaluation will be used to fully articulate the steps and the resulting benefits of utilizing this system. The evaluation of the Massachusetts ELT initiative was a goal of Massachusetts 2020 (Mass 2020), an advocacy organization, focused on the expansion and redesign of the school day. During the 2010-11 school year, 19 schools in nine districts implemented the initiative. Mass 2020, with support through a research grant from the Institute of Education Sciences (IES) within ED, assessed the implementation of ELT in these schools. Data were gathered from school administrators and teachers through interviews and surveys.

ELT developers theorize that schools that successfully implement the identified key components would provide (1) more instructional opportunities in core subjects for students; (2) increased enrichment opportunities that engage students in learning; and (3) increased opportunities for educators to plan, collaborate, and participate in professional development opportunities. Ultimately, improvements in these short-term outcomes lead to increased student achievement.
ELT developers also identify six key components that are necessary precursors to the high-quality implementation of ELT, as evidenced in the logic model (see Figure 1). The model represents the relationship between the delivery of core activities and inputs and the intended outcomes. Delivery of these six inputs and activities are facilitated by individuals schools. They include: 1) school-wide dissemination of instructional focus and assessment analysis, 2) scheduling of instructional time and academic supports, 3) enrichment activities and specials, 4) teacher professional development, 5) principal support, and 6) ELT support. The system of measurement, analysis, and reporting presented below specifically targets the implementation of these six key components.

**Intervention / Program / Practice**

The focus of this manuscript is the measurement of fidelity of implementation – a system that measures the degree to which key components, referred to as intervention inputs, are delivered and determines if that delivery was adequate to consider the intervention successfully implemented. The system utilizes a multi-step process in which 1) a comprehensive logic mode identifies key components, 2) activities associated with full implementation are further identified thereby operationalizing each component as a set of measurable indicators for which data sources and a data collection plan is clarified, 3) a scoring range for each indicator is calculated and a summary range across indicators for each key component is ultimately produced, and finally 4) a threshold or “cutoff” score that constitutes “adequate” implementation for each component is established. Ultimately, this system allows evaluators of any educational intervention to provide a comprehensive picture of the intervention and to conclude if particular components of the intervention were implemented with fidelity. Additional conclusions can be made regarding the overall level of implementation across components and across all participants of the intervention at various levels of focus. Analysis of the findings enables evaluators to answer the questions: Was the intervention delivered? And, was it delivered with enough fidelity to warrant further analysis of intervention effects.

As developers specify the intervention model and evaluators measure the fidelity with which key components of the intervention are delivered, both are faced with a variety of questions relating to measuring implementation at different levels of delivery (e.g., teacher vs. school vs. district), weighting components differently when calculating a determination of across-component fidelity, as well as identifying and accounting for theoretically-driven changes in implementation and associated thresholds from year to year. As the four steps of the system are demonstrated in this presentation, consideration of these challenges and alternate approaches in answering these questions will be discussed.

**Findings / Results**

The application of this measurement system using Mass ELT data has brought about two sets of findings. For one, we understand implementation of Mass ELT more comprehensively. Secondly, this application has enabled the developers of the system to draw conclusions about its value and contribution to the field.
What does using this system tell us about Mass ELT? Analysis of Mass ELT data indicates that the key components were not adequate implemented when applying findings to school- and sample-level thresholds set by the authors. (Note: this assumes the scoring and thresholds accurately represent developer’s original model). Each of the six components were represented by 4-7 indicators. Results show that 50% or more of indicators for three of the six components were implemented with fidelity, yet in all six cases, the established threshold across indicators were not met. Table 1 lists the indicators representing one of the six key components (i.e., school-wide dissemination of instructional focus and assessment analysis). The table includes definitions of each indicator, source of data, collection schedule, the school-level threshold per indicator, the rolled-up threshold across the entire sample for each indicator, fidelity results, a definitive determination of whether each indicator was implemented with fidelity, and lastly, a component-level determination of fidelity. Individually, each component was not implemented with fidelity. Overall, across components, the program was not implemented with fidelity, giving us reason to believe the supports and resources essential for ELT implementation were not put in place to the extent necessary to expect widespread success with short- and long-term outcomes (see Table 2 for detailed across-component, sample-level determination of fidelity).

How does this system benefit the field of educational evaluation? The method described and applied above systematizes implementation fidelity data measurement, analysis, and reporting across potentially divergent interventions and provides needed guidance to evaluators on what to measure, why, and how to use implementation data in a meaningful way. There are existing guides available on what to measure (see Century, Rudnick, & Freeman, 2010), but this system provides direction on how to synthesize and report implementation data in a manner that provides large-scale evaluations across multiple studies a valuable means to report implementation findings systematically and consistently. This system holds promise for complicated initiatives like i3 with goals of determining intervention effects. No two i3 interventions are the same, yet this system provides a common metric with which to measure and report intervention delivery.

Conclusions

The application of this system of measurement, analysis, and reporting with Mass ELT data provides evidence that the utilization of this method benefits research targeting individual interventions, as well as evaluations across multiple interventions. The method can be applied to any number of interventions regardless of domain, target population, and even method of delivery, as it is guided by the program developer’s proposed logic model and can, therefore, accommodate a wide range of interventions. Moreover, the systematic process of calculating thresholds for indicators and components that ultimately results in a dichotomous rating of fidelity (i.e., implemented with or not with fidelity) across components for the entire study sample, however, illustrates that this system can be used to synthesize implementation results, in aggregate, across multiple interventions.
References


Figure 1. Mass ELT Logic Model

- **Inputs**
  - School-Wide dissemination of instructional focus and assessment analysis
  - Scheduling of instructional time and academic supports
  - Enrichment activities and specials
  - Teacher professional development
  - Principal support
  - ELT support

- **Outputs**
  - Instructional opportunities in core subjects for students
  - Enrichment opportunities that engage students in
  - Opportunities for educators to plan, collaborate, and participate in PD

- **Short-term Outcomes**
  - Engagement
  - Communication and problem-solving skills
  - Positive behavior

- **Long-term Outcomes**
  - Academic achievement
  - Positive attitudes about teaching
  - Involvement and compensation
  - Opportunities to develop meaningful relationships with students and peer educators
<table>
<thead>
<tr>
<th>Key Elements of Component</th>
<th>Operational Definition for Indicator</th>
<th>Data Sources</th>
<th>Data Collection Schedule</th>
<th>Fidelity Scoring System (Teacher)</th>
<th>Fidelity Scoring System (Sample)</th>
<th>Results: (Sample)</th>
<th>With fidelity? [Yes/No]</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Construct 1: School-wide dissemination of instructional focus and assessment analysis</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Academic focus</strong></td>
<td>School-wide Academic Focus (SWAF) is identified by teachers and principal</td>
<td>Teacher survey; principal survey</td>
<td>Spring</td>
<td>N/A</td>
<td>0 = 75% and fewer principal/teachers agreement of identification; 1 = &gt;75% principal/teacher agreement</td>
<td>With fidelity = &gt;75% scored &quot;1&quot;; not with fidelity = 75% or below scored &quot;1&quot;</td>
<td>5 out of 18 (28%) scored 1</td>
</tr>
<tr>
<td><strong>Dissemination of focus</strong></td>
<td>Academic focus is posted publically or disseminated</td>
<td>Teacher survey; principal survey</td>
<td>Spring</td>
<td>N/A</td>
<td>0 = 75% or fewer principal/teacher agreement of posting; 1 = &gt;75% principal/teacher agreement</td>
<td>With fidelity = &gt;75% scored &quot;1&quot;; not with fidelity = 75% or below scored &quot;1&quot;</td>
<td>13 out of 18 (72%) scored &quot;1&quot;</td>
</tr>
<tr>
<td><strong>Academic-based instructional practices</strong></td>
<td>SWAF influence of instructional practices</td>
<td>Teacher survey</td>
<td>Spring</td>
<td>N/A</td>
<td>0 = 50% or fewer teachers report influence; 1 = 51-75% teachers report; 2 = &gt;75% teachers report</td>
<td>With fidelity = &gt;75% scored &quot;1&quot; or &quot;2&quot;; not with fidelity = 75% or below scored &quot;1&quot; or &quot;2&quot;</td>
<td>18 out of 18 (100%) schools scored &quot;1&quot; or &quot;2&quot;</td>
</tr>
<tr>
<td><strong>Data analysis system</strong></td>
<td>Progress monitoring / instructional adjustment through data analysis</td>
<td>Teacher survey</td>
<td>Spring</td>
<td>N/A</td>
<td>0 = 75% or fewer teachers report data are used to monitor student progress and to adjust instructional practices; 1 = &gt;75% teachers report</td>
<td>With fidelity - &gt;75% scored &quot;1&quot;; not with fidelity = 75% or below scored &quot;1&quot;</td>
<td>14 out of 18 (78%) schools scored &quot;1&quot;</td>
</tr>
<tr>
<td><strong>Dissemination of focus and assessment analysis (at sample-level)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2: Sample-level determination of fidelity: Aggregate results across components, across schools

<table>
<thead>
<tr>
<th>Key Component</th>
<th>Focus</th>
<th>Threshold</th>
<th>Results</th>
<th>Weight</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Dissemination of focus and assessment analysis</td>
<td>With fidelity = at least 75% of indicators are implemented with fidelity; not with fidelity = 75% or below implemented with fidelity</td>
<td>2 out of 4 (50%) indicators were implemented with fidelity</td>
<td>16.7% (1/6)</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>Implementation of schedule and supports</td>
<td>With fidelity = at least 75% of indicators are implemented with fidelity; not with fidelity = 75% or below implemented with fidelity</td>
<td>3 out of 6 (50%) indicators were implemented with fidelity</td>
<td>16.7% (1/6)</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>Implementation of Enrichment activities and specials</td>
<td>With fidelity = at least 75% of indicators are implemented with fidelity; not with fidelity = 75% or below implemented with fidelity</td>
<td>5 out of 7 (71%) indicators were implemented with fidelity</td>
<td>16.7% (1/6)</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>Implementation of teacher professional development</td>
<td>With fidelity = at least 75% of indicators are implemented with fidelity; not with fidelity = 75% or below implemented with fidelity</td>
<td>2 out of 5 (40%) indicators were implemented with fidelity</td>
<td>16.7% (1/6)</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>Principal support</td>
<td>With fidelity = at least 75% of indicators are implemented with fidelity; not with fidelity = 75% or below implemented with fidelity</td>
<td>1 out of 4 (25%) indicators were implemented with fidelity</td>
<td>16.7% (1/6)</td>
<td>No</td>
</tr>
<tr>
<td>6</td>
<td>ELT support</td>
<td>With fidelity = at least 75% of indicators are implemented with fidelity; not with fidelity = 75% or below implemented with fidelity</td>
<td>2 out of 5 (40%) indicators were implemented with fidelity</td>
<td>16.7% (1/6)</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td></td>
<td>With fidelity = at least 50% of components are implemented with fidelity; not with fidelity = 50% or below implemented with fidelity</td>
<td>0 out of 5 (0%) components were implemented with fidelity</td>
<td></td>
<td>No</td>
</tr>
</tbody>
</table>