Measuring effectiveness of school-based individualized interventions

Spring 2016 Update

Scott Woitaszewski, Ph.D., NCSP
Program Director & Professor of School Psychology
Licensed School Psychologist
INTERVENTION INTEGRITY

Many school-based helping professionals utilize a “consult and hope” philosophy. In such cases, a student intervention plan is developed, recommendations are made to teachers, parents, or other caregivers and best wishes for success are offered. The helping professional may “swoop in and swoop out” without insufficient attention given to proper measurement of intervention process, growth, or outcomes.

Intervention Integrity is the term used to reflect the degree to which a planned intervention was implemented as designed (Windram, 2009). Other terms such as “Intervention Fidelity” and “Treatment Integrity” may be used for the same concept, but “Intervention Integrity” will be used for the purposes of this document. In a survey of nationally certified school psychologists, only 11% claimed to “always” assess intervention integrity during consultation and no integrity records were found in teaming procedures 67% of the time (Cochrane & Laux, 2008). More recently, with data-based decision making becoming a more noted emphasis, school-based helping professionals are more inclined to consider a systematic approach to determining intervention effectiveness, starting with evaluating intervention integrity.

When considering intervention integrity, an intervention team must ask specifically, “Is the intervention being done at all?” “Is the intervention being done well?” and “What are the reasons for progress or lack of progress with an intervention?” A clear and systematic intervention integrity measurement process will help answer those questions.

Intervention Integrity measurement is a necessary component of all interventions for students in schools. Without measurement of integrity, questionable plan adherence and unknowns about the true effectiveness of an intervention approach can result. If intervention integrity is low, the outcomes are by nature, unclear. When an evidence-based intervention is used and intervention integrity is high, a positive outcome can be expected. See Figure 1 for a visual representation of this process.

Figure 1
What are the reasons for intervention progress or lack of progress?
(Adapted from Gresham, 2007)

<table>
<thead>
<tr>
<th></th>
<th>Evidence-Based Intervention</th>
<th>Unknown/Unclear Intervention Support</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HIGH Intervention Integrity</strong></td>
<td>Positive intervention effectiveness expected</td>
<td>Possible positive outcome (left to chance)</td>
</tr>
<tr>
<td><strong>LOW Intervention Integrity</strong></td>
<td>Possible positive outcome (left to chance)</td>
<td>Poor outcome expected</td>
</tr>
</tbody>
</table>
Measurement of intervention integrity must be carried out with significant care and thought. The goal is not to oversee and dissect everything a teacher or interventionist does. Rather, those responsible for measuring intervention integrity must have a good working relationship with those responsible for carrying out the intervention. When the relationship is collaborative, performance feedback about the intervention process becomes more palatable. In fact, written and verbal performance feedback has been found to be extremely valuable to successful intervention implementation (Windram, 2009). To build rapport, school-based helping professionals are encouraged to consider helping implement interventions (not just consult about them) when working with a new teacher or caregiver. Data can be collected and shown to others, and the process can be modeled. The ultimate message can be “Here is how we can make your life easier!” and “I’m available to help” (Windram, 2009).

Specific Intervention Integrity Tools

There are multiple ways to measure intervention integrity. Gresham, MacMillan, Beebe-Frankenberger, and Bocian (2000) suggested schools consider using direct observations, rating scales (self- or other), permanent products, and interviews. Windram (2009) argued that direct observation is the most reliable of these methods, as options like self-reporting are generally considered to be of low reliability and should be used sparingly or not at all. The St. Croix River Education District (SCRED) in Minnesota has developed evidence-based Intervention Scripts with associated intervention integrity checklists (for observers). These scripts and integrity checklists are saved and stored for easy access. If the intervention team needs to address a case of test anxiety, for example, an evidence-based intervention script and integrity checklist may be available already. See Figure 2 for a sample intervention integrity rating scale used by an observer of a classroom reading intervention (Gresham, 2007). See Appendix A for a more general integrity checklist that could be adapted as needed. A system that measures the percentage of steps completed successfully is often most helpful.

**Figure 2: Intervention Integrity Rating Scale for a Reading Intervention**
*(Adapted from Gresham, 2007)*

<table>
<thead>
<tr>
<th>Component</th>
<th>Low integrity</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Classroom Organized</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Teacher uses scripted lessons</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Uses appropriate error correction</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>
Quick pacing

Provides multiple opportunities to respond

Provision of positive reinforcement

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>

Total = 19/24 (79%)

Interpreting Intervention Integrity Results

Gresham et al. (2000) suggested intervention teams should strive for at least an 80% level of intervention component completion. According to those authors, at 80% or better an intervention may be considered properly completed. In the rating scale example in Figure 2, the intervention would be viewed as nearing sufficient integrity (79%), according to Gresham’s guide. In contrast, the SCRED district strives to check integrity through direct observation within 48 hours and then as often as needed in order to get to 100% integrity (Windram, 2009). Performance feedback is utilized to help those implementing the intervention become more skilled. Observers may graph integrity data or provide other written feedback. As an expected component of the intervention process, the SCRED district provides regular training to all teachers and school psychologists on this process. Ongoing integrity evaluation is critical, as intervention integrity among teachers completing interventions has been found to decrease to low levels within 10 days of implementation (Hagermoster-Sanetti & Kratochwill, 2009). Given this tendency, intervention teams are encouraged to check intervention integrity periodically, even after high or 100% integrity has been determined.

The study of how intervention integrity is defined, measured, and used is still a work in progress. A level reflecting sufficient implementation has not been agreed upon. Additionally, in the examples provided here, all components were given equal weight, a common practice that can be questioned and debated. There are often unknowns about what components of each intervention are most critical to an effective outcome. For example, it is conceivable that one component of a behavioral intervention may be twice as important as some other component of the intervention. School-based helping professionals are encouraged to be flexible, giving thought to the process before, during, and after an intervention, and collecting data about the process itself, with a goal of continual improvement.
MEASURING INTERVENTION EFFECTIVENESS

Beyond assessing intervention integrity, interventions must be measured for general effectiveness. Most single-case design interventions will include one baseline “phase” and one intervention “phase.” See Figure 3 for a visual representation of this type of phased intervention.

**Figure 3**
(Designed with ChartDog, found at [www.interventioncentral.org](http://www.interventioncentral.org)).

![Graph showing Alicia’s interfering behavior over time](image)

This phased system is also known as an **AB design**, where the A represents the baseline and the B represents the intervention. Other designs (e.g., ABAB) may be considered for your intervention, but are not required. See Riley-Tilman & Burns (2009) for additional design ideas. While an AB design has limitations, it is usually considered an adequate way to measure the intervention progress for the child in question. The key question for each intervention is as follows:

“**Is there meaningful change from the baseline phase to the intervention phase?**”

In addition to assessing intervention integrity, periodic data about the behavior in question must be collected during both the baseline and intervention phases in order to answer this question. Tallying the **frequency** of an observable behavior or skill works well in this system. Keeping frequency data is usually a simple and acceptable way to monitor a student’s progress for most teachers or interventionists. In some cases, a decrease in behavioral frequency will be desired (e.g., physical aggression). In other cases, an increase in frequency will be desired (e.g., words read per minute). However, not every intervention opportunity lends itself easily to frequency counts. For example, a struggling student may have an average of one lengthy and serious “meltdowns” per week. Showing a significant reduction in the frequency may be difficult and
perhaps impractical. Nevertheless, a team may wish to reduce the number of minutes per week the student engages in the meltdown behavior. In such cases, the intervention team could measure “minutes” as opposed to “number of incidents” and set a goal of reducing minutes from 90 minutes at baseline to a more realistic level (e.g., 30 minutes or less) during the intervention phase.

In an AB design, a significant change from the baseline is needed to imply intervention effectiveness. While not clear “proof” of effectiveness, an AB design with measurement of growth between phases takes school-based helping professionals beyond the “consult and hope” method to a more reliable data-based decision making process.

**Specific Techniques for Measuring Effectiveness**

Traditionally, even when data are collected and graphed systematically, many educators have simply “eyeballed” the data to estimate intervention effectiveness. Using this method, team members simply look at the graphs and approximate effectiveness based on the visual representation of the data. Depending solely on this method can be problematic, as intervention team members may have differing ideas on what should be considered “enough change.” Some team members may expect complete elimination of an interfering behavior or full growth to grade or age level expectations. Such goals may not be realistic for some problems or some students. In many cases, the most appropriate goal will include meaningful change or growth from the baseline level through the intervention phase, even if the interfering behavioral challenge has not been completely eliminated.

Fortunately, going beyond the “eyeball” approach, more reliable methods for measuring effectiveness are available. The Effect Size (ES) or the Percentage of Non-Overlapping Data Points (PND) can be calculated through the ChartDog application, found free on-line at [http://www.interventioncentral.org](http://www.interventioncentral.org).

Effect Size (ES) can be defined as the “magnitude of the difference” between phases. It can assist school-based helping professionals estimate the “practical significance” of an intervention. Additionally, the ES helps answer important questions such as: “Can we be confident in the impact of the intervention?” or “Are the changes observed large enough or are they too small to be meaningful in the real world?” The Effect Size option should be the preferred choice for answering such questions when there is considerable variation in the baseline data (i.e., there is no clear visual trend in the data). In the example in Figure 3 (Alicia) an inconsistent baseline situation has been observed and using the ES is appropriate. Several varied ES options exist, but ChartDog calculates the No Assumptions Effect Size (NAES). According to McGill & Busse (2014), the NAES is much like a traditional effect size and can be interpreted accordingly (e.g., with Cohen’s recommendation of .80 being large, .50 being moderate, and .20 or below being small).

As an alternative to using ES, the Percentage of Non-Overlapping Data Points (PND) is appropriate when the baseline data points are relatively consistent. PND is the most common way to determine the magnitude of change for interventions (Riley-Tillman & Burns, 2009). Choosing the PND for a consistent baseline is appropriate because an inconsistent baseline (one
with obvious “peaks and valleys” among the data points) makes it more likely that baseline and intervention phases will “overlap” (i.e., data values are of a similar range within each phase), resulting in a conclusion of insignificant growth. Like the Effect Size measure, the PND is also an estimate of the change in magnitude across phases of an intervention. While not without limitations, the ES and PND measures add reliability to the decision making process.

**Using the ChartDog Graphmaker**

To create AB design graphs and to assess Effect Size (ES) or the Percentage of Non-overlapping Data (PND) points, follow steps 1 to 6, as summarized next:

1. Go to [http://www.interventioncentral.org](http://www.interventioncentral.org). In the menu bar on the right side of the opening page to find and click on *ChartDog Graph Maker*.

2. **Click the black bar showing CHART to set up the chart structure.** Most of the default settings are appropriate, but consider adding titles for the overall chart and for the vertical and horizontal axes. This can be done by clicking on *Chart Options* from a menu at the bottom of the chart. From there, type in the chart title, the X axis title and the Y axis title. See Figure 3 for examples of titles added.

3. **Click the black bar showing DATA to enter data into your chart.** Find the box showing MY PHASE. Click on the pencil icon to edit that phase title. Re-name it BASELINE and adjust dates according to your baseline phase. Next, find and click on *Create A New Phase*. Add a new phase called BEHAVIOR INTERVENTION PHASE and adjust dates according to your intervention. Doing all of this will create an AB design structure for your single-case intervention data.

4. **Add your baseline and intervention phase data.** Directly to the right of the baseline phase box, click on *Create New Entry*. From there, add your data points one by one for the baseline phase by entering in a number in each VALUE box. The value entered will be based on how you measured the intervention (e.g., frequency of the behavior, minutes, etc.). Be sure the date is correct for each entry. Repeat this process for the intervention phase (i.e., click on *Create New Entry* and proceed as above).

5. **Choose analyses to determine if a significant change has occurred.** With the goal of evaluating change from the baseline through the intervention phase, click on the black bar showing DATA ANALYSIS. Click on MEAN BY PHASE OFF to calculate the value means for each phase. This process will give you an overall sense of change from phase to phase. Next, choose PERCENTAGE OF NON-OVERLAPPING DATA (PND) or NO ASSUMPTIONS EFFECT SIZE (NAES). Your choice will be based on the discussion noted above (pages 6-7). After choosing, be sure to set up the analysis so the *Base Phase* is your baseline phase and the *Treatment Phase* is your behavior intervention phase.

6. **Interpret the results.** Chart Dog will provide you with PND or ES data results. Use the following guide for interpreting ES and PND data. *NOTE*: a negative ES simply implies a
lower intervention phase mean relative to the baseline. The same interpretive guide can be applied to both positive and negative Effect Size values.

Percent of non-overlapping data (PND):
- PND of 85%+ = Highly effective
- PND of 65 to 84% = Moderate effectiveness
- PND of 50 to 64% = Questionable effect

Effect size (ES) (Cohen, 1988):
- +/- .80 = Large Intervention Effect
- +/- .50 to .79 = Moderate Effect
- +/- .20 to .49 = Small Effect

Evaluation of Intervention Effectiveness Case Study: Alicia’s Aggression

Alicia is a 6th grader at Adams Middle School. She has recently struggled significantly with physical aggression toward peers in school, often demonstrating several occurrences of hurtful pushing or shoving each day. After participating in several Tier 1 and Tier 2 behavioral intervention approaches, the intervention team decided Alicia could benefit from a more intensive Tier 3 intervention (individualized). The intervention team developed a plan that included wraparound services, a strong system of frequent encouragement and positive reinforcement, and weekly check-ins by the school counselor. Prior to the intervention starting, Alicia’s teachers monitored instances of physical aggression by tallying frequency during each hour of school for one week (a baseline phase, see Figure 3). A total frequency count for each day was calculated for the baseline and then for the intervention phase (two additional weeks beyond the baseline). Those total daily frequency counts were entered into the value boxes in the ChartDog data entry process. Given the relatively varied baseline data that resulted, the intervention team chose to use an Effect Size (NAES) as the primary determinate of intervention effectiveness. The results are as follows:

<table>
<thead>
<tr>
<th>NAES</th>
<th>MEAN BASE</th>
<th>MEAN TREND</th>
<th>STD DEV</th>
<th>SERIES</th>
<th>BASE PHASE</th>
<th>TREATMENT PHASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.85</td>
<td>3.63</td>
<td>2.13</td>
<td>1.77</td>
<td>Series 1</td>
<td>Baseline</td>
<td>Intervention Phase</td>
</tr>
</tbody>
</table>

As can be seen in the NAES cell above, the ES for Alicia was – .85, a value that exceeds the +/- .80 value suggested by Cohen (1988) to reflect a large effect. The mean levels of each phase suggest the physical aggression had not been eliminated completely, as the mean during the intervention phase was still 2.13 incidents per day. However, the reduction in physical aggression from the baseline phase (mean = 3.63) is still notable, with the large effect size providing a powerful indicator of significant change.
While the Percentage of Non-Overlapping Data (PND) approach was not emphasized in this intervention phase (due to a highly varied baseline), Alicia’s PND is included next (for demonstration purposes only). As can be observed below, the PND was 33%, indicating that one-third of the intervention phase data points were lower than any data points from the baseline phase. With aggressive acts in mind, we know that it is good to see any data points that are below baseline levels. However, according to the PND interpretive guide, 33% is considered a low or insignificant level of change. Again, the PND was not viewed as an appropriate interpretive choice for Alicia’s data because of the highly varied baseline. In such cases, it can often be assumed the PND results will not result in conclusions of effectiveness. Instead, the ES was the appropriate choice in Alicia’s case and would be in other cases like this.

<table>
<thead>
<tr>
<th>PND</th>
<th>BASE PHASE</th>
<th>TREATMENT PHASE</th>
<th>DIRECTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>33.33</td>
<td>Baseline</td>
<td>Intervention Phase</td>
<td>DOWNWARD</td>
</tr>
</tbody>
</table>

Finally, it should be noted that, through several direct observations by a school-based mental health professional who monitored intervention implementation, the integrity of this intervention was found to be 100%. Overall, given strong intervention integrity and a large effect size, the intervention team had meaningful data supporting the likelihood of intervention effectiveness in Alicia’s case. The team decided to continue the intervention for the foreseeable future.

Cautions in Evaluating Intervention Effectiveness

The rather large difference noted between the two evaluation methods considered in Alicia’s case (ES and PND) – while using the same set of data – is clearly noted. One method (ES) leads us to concluding a large effect; the other leads to a small or insignificant effect (PND). This condition is a shortcoming of attempting to evaluate individualized interventions with some type of outcome indicator. Despite frequent use, there is no consensus on the best method for analyzing the effectiveness of single-case design interventions (McGill & Busse, 2014). The ideas and guidance offered in this document improve the process of school-based intervention evaluations but do not offer full assurance in the resulting conclusions.

References


