The Essential and Interrelated Components of Evidenced-Based IEPs

A User’s Guide

Marc D. Hauser
The Tanners’ son, William, receives special education services. William has a hard time with emotional regulation when frustrated. One of his individualized education program (IEP) goals states, “William will use coping skills when he is frustrated, on 75% of the observation periods.” Upon reviewing this goal, William’s special education teacher asks, “What, in behavioral terms, does ‘frustration’ look like? What coping skills is he using? Given that there is no baseline, why is the benchmark 75% of the time? Is this 75% of a day on average, or 75% of the days in which observations are recorded?” This goal is not educationally meaningful and cannot be reliability monitored and objectively analyzed. What can she do?

There are many books and articles on the process of writing IEP goals (Bateman & Herr, 2011; Gibb & Dyches, 2007; Peterson, Burden, & Sedaghat, 2013; Twachtman-Cullen & Twachtman-Bassett, 2011). There are also several “goal banks,” many published online (e.g., GoalBook, IEP Goals), that provide practitioners with specific examples across all content-relevant areas. In addition, there are resources that explain methods of evaluation, including scales of measurement, baselines, reliability, and statistical analyses (Chafouleas, Riley-Tillman, & Sugai, 2007; Cook, Carter, & Cote, 2014; Gunter, Callicott, Denny, & Gerber, 2014; Martin & Bateson, 2007). Despite these resources, many special educators point to the poor quality of written goals and the evidence collected, raising questions about both what to measure and how to measure it.

Many special educators point to the poor quality of written goals and the evidence collected, raising questions about both what to measure and how to measure it.

This article describes a framework that special educators can use to write and evaluate IEP goals and objectives. By following the guidelines and recommendations in this framework, special educators will find it easier to develop meaningful and measurable IEP goals and objectives. The enhanced quality of these goals and objectives will in turn facilitate collecting more reliable data related to student growth and program effectiveness.

Consider the following fairly typical IEP goal from Wrightslaw:

Using real money, Jane will be able to show how much money she has after she receives two weeks of allowance, and how much money she will have left after she buys one object, with 75 percent accuracy measured twice weekly each quarter. (Heitin, 2014)

Although this goal includes a definition, benchmark, and frequency of evaluation—the standard components commonly seen in IEPs—there are three problems. First, the description of the behavior evaluated is ambiguous, leaving too much room for error in data collection. Who determines how much money to allocate to Jane each session, the range, and the currencies? Does Jane pick the object she wishes to buy or does a staff member, and what is the range of values for the objects? Without ironing out the details of the behavior, reliable data collection is impossible.

Second, the frequency of evaluations is not specific. Saying that a behavior will be “measured twice weekly each quarter” fails to specify the interval between opportunities, the time of day, or the minimum number of evaluations within the quarter. Without such information, it is not only impossible to replicate the method but difficult to assess the validity of the data collected. Decades of studies in the cognitive sciences show that the interval between learning opportunities affects not only what is learned but how well it can be used and generalized to new contexts (for a review, see Roediger & Pyc, 2012).

Third, the benchmark refers to only a single descriptive statistic (75% accuracy) that is neither explained nor considered in light of any variability in performance. For example, imagine that Jane reaches her 75% benchmark, but during this period she also dropped to 30% and soared to 100%, with scores in the past 2 weeks averaging only 65%. Although she objectively reached her benchmark at one point in time, this single value provides an incomplete and misleading account of her competence.

The framework described here resolves these concerns by clarifying the four essential and interrelated components of IEP goals and objectives. (For stylistic reasons, I use the word goal most frequently, although the issues discussed generally apply to both goals and objectives.) In general, goals provide an overarching description of the skill or behavior targeted over the course of the year, whereas objectives are more short term, describe subcomponents of the behavior or skill, and provide more detail on terminology, benchmarks, and methodology. A goal and its objectives are typically articulated when it is helpful to specify not only...
the subcomponents of a skill but the benchmarks and timeframes for each. In contrast, the goal alone may be specified when it encapsulates a single, specific learning skill, such as achieving competence in decoding words on the Gray Oral Reading Test.

This framework uses the acronym OBSE, which stands for operationalize key terms associated with skills and behaviors, using clear and measurable definitions; benchmark performance targets by establishing appropriate statistics and timelines; scale the observations or responses in terms of quantitative or qualitative values; and evaluate the methodological details in terms of who will collect the data, when, where, and how often. All of these components are closely interconnected. Figure 1 provides an IEP template with an example of one well-worked goal and objective (in this case, the second objective would most logically focus on William’s ability to independently use coping skills when frustrated); Table 1 is a checklist to help IEP teams avoid pitfalls and implement effective solutions.

**Figure 1. IEP goal and objective template**

**Annual goal description**

William Tanner will use different coping skills (e.g., go on a walk) when he is frustrated, as evidenced by specific behaviors (e.g., verbal outbursts, head down). The overall goal is to reduce the intensity of frustration, and its behavioral manifestations, by enabling William to independently and effectively use coping skills.

**Objective 1**

**Operationalize** The aim is for staff to help William use different coping skills when either he or the staff recognize the behavioral indicators of frustration. The primary indicators are William saying “I’m frustrated”, putting his head down, fidgeting with an object (e.g., pen), or getting up and pacing while making sounds (e.g., groaning). When staff see these indicators, they should immediately offer William one of the following coping skills: a walk outside or in the hallway, the use of a settling area where he can draw or read a book. William should immediately (within 2 minutes) accept and use the skill, without arguing or turning away from staff.

**Benchmark** By the end of the first quarter, William will accept and then use a staff-offered coping skill on an average of 80% of all opportunities. In addition, during the last 5 observation days of the quarter, William must accept 100% of all staff-offered coping skills. The reason for demanding this high level of coping skill use is to enable a successful transition to Objective-2, focused on independent use of coping skills.

**Scale** For every observation where William shows frustration and a staff member offers a coping skill, use a 3-point scale for scoring. Score 0 if William rejects (turning away from staff or arguing) the coping skill after 3 or more prompts and more than 5 minutes of time elapses from first offer; Score 1 if he accepts (without turning away or arguing) after 1-2 prompts and starts using the skill 2-5 minutes after the initial offer; Score 2 if he accepts without prompts, argument or turning away within 2 minutes.

**Evaluate** All staff in the classroom involved in evaluating William will first review the types of coping skills offered as well as the behavioral indicators of frustration. For the first two weeks of evaluation, an attempt will be made to have 2 or more staff evaluate each observation of frustration, with one staff member then being responsible for offering the coping skill. The other staff will observe the interaction and score William’s response on the 3-point scale. This initial 2-week period will be used to refine the method and obtain agreement among staff. At the end of each day record the number of points earned on the 3-point scale for each individual episode of frustration, and divide this by the number of points that could be earned (2 points times the number of opportunities where frustration observed and coping skill offered). This generates a percentage for each day.

*Note.* IEP = individualized education program.
Operationalize

The heart of most IEP goals is a definition of the skill or behavior that requires attention—the operationalization (Bateman & Herr, 2011; Peterson et al., 2013). It is essential to precisely define each term. A simple litmus test for precision is to answer two questions:

1. Given the terms used, do I know how to measure the behaviors or skills?
2. On the basis of how I understand the terms, can I imagine that other evaluators might have different interpretations of how or what to assess?

If your answers are no and yes, then the goal does not have well-defined terms. To illustrate, consider a seemingly straightforward goal: “Reduce the number of times Fred is out of his chair.” Presumably, the answer to Question 1 of the litmus test is yes—score the number of times Fred is out of his chair. Question 2 is problematic because staff could provide different definitions for out of his chair. Some staff may consider Fred out of his chair as soon as he breaks contact with the seat of the chair, whereas others may be more lenient and consider him out of his chair only if there is no contact at all. If such variation exists, then there will also be variation in the scores, which will create a fundamental problem of interpretation. Is the variation observed in Fred’s out-of-chair behavior due to Fred, to the staff observing him, or both? Problems like this are magnified for social-emotional goals and objectives.

Consider William’s goal: “William will use coping skills when frustrated.” This type of goal fails both litmus test questions: coping skills are not defined and frustration is not expressed in measurable behavioral terms. In the absence of definitions, evaluators are likely to provide very different interpretations. To satisfy the litmus test, spell out the specifics of coping skills (e.g., doodles, uses a stress ball) and the behaviors observed when frustrated (e.g., fidgets, makes huffing sounds). Figure 1 fully develops this goal and its objectives using the OBSE framework. Table 2 provides examples of unmeasurable goals and objectives, the reasons they are unmeasurable, and a revision that is measurable.

---

**Table 1. Checklist for IEP Goals or Objectives**

<table>
<thead>
<tr>
<th>Component</th>
<th>Activities</th>
</tr>
</thead>
</table>
| Operationalize | - Write the goal or objective like a user’s manual—tell the staff what behavior or skill to look for and, thus, what to evaluate.  
- Check for ambiguous words (terms that could be interpreted in different ways).  
- Explain all emotional, motivational, or mental state terms (e.g., belief, intend) by providing observable behaviors.  
- Each goal or objective should refer to only one measurable behavior or skill. |
| Benchmark | - Determine whether baseline or published data are available, and, if so, assess quality. Baselines are essential for setting benchmarks.  
- Create different descriptive statistics for each performance target, including information on the child’s variability over time.  
- Justify the use of target values by assessing whether higher or lower values are required to advance to the next step in the process, whether another objective, grade, or school. |
| Scale | - Determine whether you are assessing the quantity, quality, or both for a behavior or skill.  
- Determine whether staff members have the time to collect the relevant data given the scale assigned. For example, if counts are required, do they have time to accurately count given other demands on their time?  
- Define the terms of your scale so that staff members understand how to score performance or the expression of a behavior.  
- If you use a categorical scale, make sure the options are balanced, including negative and positive values along with one that is neutral. |
| Evaluate | - Identify who will be carrying out the evaluations. If two or more staff, are they all trained, aware of OBSE?  
- Determine when the evaluations will be carried out. Are the evaluations spread out through the day, during different days of the week, and in different contexts to capture potential sources of variation?  
- Determine what the context is for evaluation. Is the evaluation carried out in a group setting, while the student is alone, in a classroom or outside, etc.?  
- Determine how the evaluations will be collected. Is the evaluation recorded on video and transcribed later, written into a check sheet, or entered into a digital database? |
At a minimum, deciding on and evaluating benchmarks depends on two broad pieces of evidence which may or may not be known: (a) baseline information on the behaviors or skills written into the IEP and (b) data on children with similar challenges or disabilities, either from the published literature (Lloyd, Ledford, & Gast, 2014) or from within a school over several years. Even when such information is available, it is important to assess its quality and relevance. For example, if baseline information was collected on the child, is there a description of how many times the student was evaluated, start and end dates, and methodology? If the baseline is presented as a single value or score, is there information on the variability leading up to this summative score? For example, and as illustrated in Figure 2, an average score of 5 out of 12 could be generated from a student who consistently scores between 4 and 6 over several weeks or could be generated from a student whose scores consistently rise or fall from 2 to 12. These different patterns would dictate different approaches to intervention, perhaps backing up to simpler material if the scores have been falling or moving ahead to new material if they have been steadily rising. When

<table>
<thead>
<tr>
<th>Content area</th>
<th>Unmeasurable operationalization</th>
<th>Reason for problem</th>
<th>Measurable operationalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Communication</td>
<td>Will express his or her opinion or idea politely and independently in group discussion after listening to other classmates’ ideas or opinions, at least four out of five times with 80% accuracy.</td>
<td>What does <em>politely</em> mean? Is it the absence of name-calling or a harsh tone, or the use of supportive expressions (e.g., “I can see your point, but I believe . . .”), or all three? Is it polite if the student just states his or her opinion or idea, without reference to the others in the group? How do you assess accuracy in a percentage when you are evaluating the content of a comment? What is <em>accuracy</em>?</td>
<td>After listening to a discussion among classmates, student will raise his or her hand, and when called upon, will first acknowledge or reiterate what others have said, without criticism, and then state, in a calm voice (nonaggressive tone) his or her own opinion or idea. Evaluation of the student’s response will focus on both the content of what he or she says, and how he or she says it. The goal is for student to use both the appropriate content and tone on four out of five opportunities.</td>
</tr>
<tr>
<td>English language arts</td>
<td>When presented with letters, numbers, or words, will match with 90% accuracy on four of five trials.</td>
<td>Which letters, numbers or words? Familiar and unfamiliar? Does “matching the letter” entail correct upper- or lowercase usage? Does matching a word entail all letters of the word, in the same order? If letters, numbers, and words are presented, the accuracy score, which is summative, cannot pull out differences between these three items (e.g., better on numbers than letters). The method of presentation is also ambiguous.</td>
<td>When presented with 10 familiar lowercase letters, one at a time, student will copy the letter with an overall score of 90% accuracy on four out of five opportunities. Each opportunity should present a different set of letters in a different order. Objectives 2 and 3 would follow a parallel structure, focused on numbers and words, respectively.</td>
</tr>
<tr>
<td>Motor skills</td>
<td>Will cut with scissors.</td>
<td>Will cut what kind of material, with what kind of pattern, and what length?</td>
<td>Will use scissors to cut along an 11-inch straight line placed at the center of a piece of white, 8.5 × 11-inch paper.</td>
</tr>
</tbody>
</table>
examining data available on similar students, who tested the children, when, and where? Were the children evaluated in a laboratory? At school? At home? Were they evaluated by unfamiliar or familiar testers? If the evaluation environment is different from the everyday school setting, then the ability to generalize is more limited. In this situation, educators should gather their own baseline for comparison purposes, attempting to match, as much as possible, the available methodologies.

Irrespective of what kind of information you have access to, it is important to consider different descriptive statistics and the reasons for particular benchmarks. For example, many IEP benchmarks are written with a single target value (e.g., 75%), often without justification. This approach is insufficient because it fails to take into account either the variability of the child’s behavior, the variability of the evaluation process, or simple common sense. As alluded to earlier, if a child scores 75% on average, but is sometimes well above and well below this score, then the variability itself may be of concern. In addition, if the goal focuses on safety to self or others (e.g., crossing the street, intense aggression), then the benchmark might be set with no tolerance for variation (e.g., 100% safe crossings, zero aggressions).

As a guideline, performance targets associated with IEP benchmarks should incorporate different descriptive statistics, including percentages, averages, sums, and measures of variability overall and during different time periods (see Table 1). Benchmarks written in this way specify a target mean (e.g., 80%) as well as a tolerable range of variation (e.g., no score lower than 70% within the past 3 weeks of evaluation; at least 10% of the scores between 90 and 100). Each statistic should be justified in terms of the overall aims of the goal. For some goals, such as those that relate to the well-being of self and others, it may be important to aim for low variability and a narrow range of scores. For example, an average of two aggressive actions per day and no more than four in 1 day may be appropriate. For other behaviors or skills, a higher average may be necessary to guarantee a smooth transition to a class that is at a particular grade level in the curriculum. For example, an appropriate math score may be 90% or better with no scores lower than 80%. Sensitivity to existing variation, as well as its causes, will provide not only a better sense of student growth but also a more informative assessment to guide interventions.

S: Scale

The measurement scale depends on what is being assessed, how it is defined, the evaluation methods, and the targets for student growth. To show the interrelatedness of these issues, consider a goal focused on decreasing a child’s aggression, defined as physical contact with a peer using any body part or an object, including pinching, biting, kicking, or throwing a rock at someone. A numerical scale would be appropriate if the aim is to record the number of aggressive acts per day. Although straightforward, this scale ignores intensity differences; pinching is quite different from biting. If the aim is to reduce both the frequency and intensity of aggression, then it may be advisable to clearly define and score the frequency of mild, moderate, and intense aggression. Adding on this additional dimension depends, in part, on the ability of staff to track all of this information, a point that highlights the relationship between the design of a measurement scale and the details of collecting data.

Considerable work has been devoted to the validity of different measurement scales, both within and outside of education (Cicchetti et al., 2006; Streiner, Norman, & Cairney, 2014). Table 1 provides a checklist of considerations. The scale used cannot be developed without consideration of the data collection methods. Consider William’s goal of using coping skills when frustrated. The IEP states that evaluation for this goal will be based on a categorical scale with five options: never, rarely, sometimes, often, or always. Each option is defined in terms of the percentage of occasions in which

Sensitivity to existing variation, as well as its causes, will provide not only a better sense of student growth but also a more informative assessment to guide interventions.
William uses coping skills. The aim, however, is for all teachers and the primary adjustment counselor to provide data. This is valuable in terms of the breadth of evaluators and thus contexts but poses a problem if some see William less often than others. For example, it is quite likely that teachers will see William every day, whereas his counselor may see him only once or twice a week. Using the never-to-always scale would raise problems as the counselor’s sample is more limited than that of the teachers. A score of rarely would mean different things but be given comparable weighting in the data. Thus, scales of measurement must be calibrated to the frequency and contexts in which different evaluators assess students. In this case, a solution would be to use only the teachers’ data or make sure that the counselor sees him as frequently.

E: Evaluate
Either before or after you have operationalized your goals and described your benchmarks and scales, you will need to design your methods for evaluating the data about a student’s behavior or skills. Your evaluation methods dictate what evidence is possible and what how long, for how many days, and with how much time between observations? Each of these factors could significantly affect the evidence collected. For example, some students may not perform as well on a skills-based task when assessed in the early morning if they are tired or anxious about the day. Ideally, data collection should be distributed throughout a day in order to capture a representative picture of performance or behavioral expression. If that is not possible, then try to set up a schedule so an equal number of observations occur in a morning and afternoon block. If data can be gathered only during a certain time of day, then discussions of performance must be restricted to this time of day. Where dictates the setting or context for data collection. As with decisions regarding when, it is important to think through the consequences of where a behavior is monitored, as students are often affected by the physical structure of the space (e.g., indoors vs. outdoors), how many teachers and students are nearby, and any distractors that might arise during an activity. Although it may be the case that skills-based data collection occurs within the student’s classroom, variation within this setting can arise due to the number of students in the class, where the student sits at the time, and perhaps uncontrollable noises that arise during a school day. As with who, it is important to control for these factors as much as possible. If not possible, then take note of them so that they can be examined in relationship to performance.

Associated with each of the three Ws are additional methodological demands focused on how. For example, whether a school uses one or more staff, it is necessary to establish criteria for assessing and quantifying reliability (Hallgren, 2012). When two or more staff are involved, do they provide consistent assessments, using the same criteria for scoring? How is consistency established, not only at the start of a sampling period but throughout the data collection period? As individuals learn more about a student, how will checking for assessment drift be handled? How are data collected, entered into a system of analysis, and checked for inaccuracies?

Providing clear methodological descriptions prior to collecting data will help ensure high-quality evidence. For example, to establish interstaff reliability, a school might start by providing all individuals with a rubric or document that spells out definitions for key terms and lays out the core steps in data collection. Next, two or more staff score the same student or students at the same time, but independently. Last, compare the scores and review any differences. If there are relatively few differences (e.g., <10%), then the method is reliable. If there are significant differences among the raters, then there are at least two corrective measures. First, review the details of the scale and evaluation methods to make sure there is agreement. Second, reevaluate a new group of students with multiple staff at the same time, but have staff discuss the first few scores before scoring independently. By discussing differences, staff can refine the scoring procedure and, thus, improve reliability.

Conclusion
The OBSE framework outlined here is designed to help practitioners see the importance of each of the four essential components of an IEP goal as well as their interrelatedness. Equipped with such knowledge, practitioners will be able to enhance the quality of their services by providing more reliable and meaningful data to assess the effect of their work on student progress. Adopting this framework represents a shift in approach, one that requires professional development, teamwork,
and an appreciation of the significance of high-quality data in guiding the development, implementation, and evaluation of strong IEP goals (Blackwell & Rossetti, 2014).

Professional development will help educators understand the concepts associated with the four components and their interrelatedness while explicitly contributing to implementation (Doren, Flannery, Lombardi, & Kato, 2013). For example, rather than attempting to implement the full framework all at once, it is often useful to start by focusing on operationalizing terms within each goal and objective. It is important to provide feedback, making sure that all terms are clearly defined, with all emotional and cognitive statements expressed in measurable behavioral terms (see Tables 1 and 2). Once this skill set is firmly in place, the professional development provider can assist in the development of benchmarks while setting up the discussion of scales and evaluation methods. Last, the professional development provider can help educators assess the feasibility of different data collection and evaluation approaches given the many other demands on their time. After all, data collection is not the educator’s only responsibility! Ultimately, all goals and objectives must be written in such a way that they not only capture the challenges each student faces but also articulate a method of evaluation that is feasible given the educator’s schedule and the resources available for assessment. Such discussions and support strengthen IEP goal development and implementation.

The team members involved in creating and implementing IEP goals must work together to find a balance between the best possible evidence and the reality of a school day (Talbott, Mayrowetz, Maggin, & Tozer, 2016). Although it may be ideal to observe and record a behavior every day, at different times of day, and using different evaluators with high interstaff reliability, the reality of the school day may not permit such methods. A compromise may therefore be necessary because evaluators have other responsibilities that permit observing or testing only every other day and only during the morning. Teamwork is essential to realizing high-quality evidence, but it is also essential in terms of identifying constraints. Thus, during the creation of an IEP or evaluating the merits of an existing IEP, it is essential to check that the desired evidence-based demands on evaluators are achievable.

In many ways, professional development and teamwork depend upon evaluators who believe that high-quality evidence is essential to helping students learn. Without an interest and passion to identify, develop, and measure meaningful IEP goals, the process has little hope of getting off the ground. One way to help educators see the value of their work is to use baseline information on each goal to establish targets for short- and long-term growth. For example, if a student’s goal entails remaining in the classroom for the entire period without a student- or staff-initiated break, baseline data establish not only what the student can presently achieve (e.g., 20 minutes out of 45 minutes) but what a desired target or benchmark would be (e.g., an increase to 30 minutes by the end of the first quarter). By keeping an eye on growth targets while collecting and visualizing the data, educators will see the effect of their work on individuals. It is this influence that motivates every special educator (Cook & Cook, 2013).

In sum, special educators are tasked with the remarkably challenging task of improving the outcomes of students with disabilities. An essential tool for facilitating this task is the IEP. Special educators who operate under clear goals and collect meaningful and reliable evidence to inform instruction are better situated to meet this task. The framework set out here will take educators closer to developing strong, meaningful IEP goals and collecting insightful and reliable evidence.

References


Marc D. Hauser. President, Risk-Eraser, West Falmouth, MA.

Address correspondence concerning this article to Dr. Marc D. Hauser, Risk-Eraser, P.O. Box 376, West Falmouth, MA 02574 (e-mail: marc@risk-eraser.com).

TEACHING Exceptional Children, Vol. 49, No. 6, pp. 420–428. Copyright 2017 The Author(s).