
Center for Research and Reform in Education (CRRE)
Johns Hopkins University

Prepared for Digital Promise, which is working in partnership with the Education Industry Association
Evaluating Evidence for Ed-Tech Product Effectiveness:
Guidelines for School Districts

Educational technology products offer potentially effective means of supporting teaching and learning in K-12 classrooms. But for any given instructional need, there are likely to be numerous product options available for purchasing. What can school districts do to help ensure that good selections are made? In a recent, comprehensive study of ed-tech product procurement\(^1\), the processes used by the consumers (school district stakeholders) and sellers (product developers and providers) were viewed along a continuum of the following five “Action Points”:

I. Allotment of Funding  
II. Assessment of Needs  
III. Discovery of Ed-Tech Products  
IV. Evaluation of Products  
V. Acquisition of Selected Products

As in buying everyday consumer products, such as cars, furniture, and washing machines, decisions certainly depend on the available budget and personal needs. In the case of ed-tech products, the latter are determined by “needs assessments,” which should specify where instructional support is needed (e.g., in fourth-grade math), how it would be used (e.g., as a supplement for regular instruction), and by whom (e.g., all students at individual performance levels). From the needs assessment, focus shifts to” discovering” which products are in the marketplace that can directly support the identified needs.

The discovery process will undoubtedly yield at least one and, frequently, several solutions that potentially “fill the bill.” Of course, there will be many factors to consider in evaluating these products. Some may have features that meet a school’s requirements better than others (e.g., a parent participation component), some may be more economical, and some may be more compatible with existing resources and space. But, arguably, the most important factor is that some may be more effective and be easier to implement than others! This consideration brings us to the focus of these Guidelines—How can school districts evaluate evidence of effectiveness?

Types of Evidence

In our study of ed-tech procurement, school district stakeholders reported relying the most on four different categories of evidence to determine product effectiveness:

---

• Recommendations from other school district peers or consultants
• Recommendations from end-users (teachers and principals)
• Findings from “pilot tryouts”
• Evidence from “rigorous” studies

In contrast, reliance on pitches from sales representativeness or evidence from informal studies was much less common.

While we orient the Guidelines below to interpreting “evaluation studies” of product effectiveness, we likewise recognize the value of peer and end-user qualitative feedback. Peers and end-users may lack concrete, scientific evidence to support particular products, but they certainly can offer thoughtful firsthand impressions of how the products were used in settings similar to those your district has targeted for ed-tech support. Specifically, a superintendent or technology director may provide information regarding the degree to which a particular product was easy to implement, was liked by students, parents, and teachers, and appeared beneficial for teaching and learning. Teachers and principals from the peer districts can also offer judgments of product utility and quality based on pilot reviews, try-outs, or demonstrations.2

**What Makes It “Rigorous”?**

When we asked school district stakeholders in the procurement study the degree to which they relied on “rigorous evidence” in selecting products, 54% of the superintendents and 58% of the principals surveyed gave ratings of “more than moderately” or “extensively.” Interviews with these same participant groups, however, indicated that rigor was defined in different ways, often in terms of who obtained the evidence—an independent evaluator or the developer.

For present purposes, it is important to establish a common and more formal definition of rigor. For evidence to be rigorous, it must be **reliable** and **valid**. While these constructs themselves have different meanings and technicalities, let’s try to keep things simple here:

- **Reliability**: The methods of the study are:
  - clearly delineated,
  - and if replicated,
    - are likely to produce consistent results, barring some substantive change in conditions, product features, or participants.

- **Validity**: The methods of the study seem likely to demonstrate the **true** impacts of the product with representative users and conditions.

---

2 District respondents in the procurement study described such activities as mostly informal, brief demos or tryouts as opposed to structured pilot or “short-cycle” studies.
These two properties probably appear somewhat similar. In fact, reliability is a necessary attribute of validity. That is, if the evaluation methods used are not well structured and, depending on which way the wind is blowing, prone to yield variable findings, validity certainly would be compromised. But reliability by itself does not guarantee validity. Consider, for example, the following scenario:

A group of 20 teachers agree to try out in their classes an ed-tech tutorial program in mathematics. After ten weeks, 8 of the teachers are continuing to use it, while 12 have dropped out. User data and impressions of effects are obtained from the former (still active) group to judge product quality. The findings, overall, are quite positive.

If the above study were repeated, the results might well be the same (i.e., quite positive). The evidence therefore seems reliable, but is it valid? Seemingly not, since only the group that persisted with the product evaluated it. Along the same lines, recently one of us heard this announcement from the flight attendant as the plane was landing: “If you feel that we provided excellent service, be sure to take the survey on the following website…” Apparently, those who perceived the service as only inadequate to good need not bother!

<table>
<thead>
<tr>
<th>Key Questions:</th>
</tr>
</thead>
<tbody>
<tr>
<td>• If the evaluation were repeated, are the results likely to be consistent (reliable)?</td>
</tr>
<tr>
<td>• If the results are consistent, do they also seem accurate, fair, and unbiased (validity)?</td>
</tr>
</tbody>
</table>

Directly related to the key reliability and validity questions are characteristics of the studies of product effectiveness. Table 1 previews our later discussion of what each characteristic entails and might be considered in evaluating available evidence.
Table 1. Characteristics of Evaluation Studies of Ed-Tech Products

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Who conducted the study?</td>
<td>A third-party (independent) evaluation has higher credibility than an internal (in-house) study.</td>
</tr>
<tr>
<td>2. What was the study design?</td>
<td>Rigorous designs that minimize error in measuring and interpreting product quality increase confidence in making judgments.</td>
</tr>
<tr>
<td>3. What were the conditions of the study?</td>
<td>Realistic conditions that mirror real-life product use make results more meaningful and higher in “external validity.”</td>
</tr>
<tr>
<td>4. How recent is the study?</td>
<td>The more current the study, the more relevant the evidence will be to present conditions (e.g., Common Core State Standards) facing teachers and students).</td>
</tr>
<tr>
<td>5. Who participated in the study?</td>
<td>Representative and unbiased sampling increases the generalizability (applicability) of the evidence to the broader population.</td>
</tr>
<tr>
<td>6. What were the outcome measures?</td>
<td>Relevant and unbiased measures of product effectiveness for student engagement, teacher satisfaction, and achievement outcomes increase confidence in the evidence reported.</td>
</tr>
</tbody>
</table>

**Identifying the Source**

While it is reasonable for providers to present evidence supporting their products, consumers need to know who conducted the studies. Was it an independent evaluator unaffected by the product’s success or an individual or group affiliated in some way with the provider or the product? Clearly, we will view evidence more confidently to the degree that we can rule-out that self-interest or bias (intentional or not) influenced the study. In this regard, some factors to keep in mind while considering the source of evidence are:

- In-house studies rarely if ever report negative findings on major outcomes.
- Providers rarely if ever release non-supportive evidence (unless part of a larger, more evidence portfolio).
- Evaluators who are completely independent of the providers (i.e., are supported by external grants and not by the provider) are the most credible sources. Only a small percentage of ed-tech products are evaluated via external funding.
- More commonly, providers commission evaluators to conduct studies of product effectiveness. Credibility is stronger if:
  - the evaluators are from reputable, well-known organization (universities, national or regional research centers, established research firms, etc.).
  - the evaluators are otherwise independent of the provider, aside from being contracted by the provider.
the evaluators deliver reports that appear fair and unbiased, noting weaknesses or negative findings where they occur, rather than only strengths.

- Because studies published in mainstream educational research journals have undergone peer review, their credibility increases over less formal sources.

**Design Essentials**

Regardless of who conducts the study, the credibility of evidence is directly and strongly influenced by the evaluation design. While many numerous and complex design options exist, practitioners and providers can be sufficiently prepared by being knowledgeable about several basic categories. These are described in Table 2 and below. Keep in mind that the categories described are quite broad. That is, within each, the studies included can vary in size, complexity, and rigor. So, judging evidence requires a bit more scrutiny than just knowing that the category is “Experimental,” “Case Study,” or “Pilot-Short Cycle”.

Table 2. Basic Types of Evaluation Designs

<table>
<thead>
<tr>
<th>Design</th>
<th>Characteristics</th>
</tr>
</thead>
</table>
| Experimental Studies             | *Compares “program” group to a control group on one or more educational outcomes (e.g., student achievement, interest, lesson completion time, etc.)  
*Uses systematic and rigorous methodology  
*Can use random assignment or self-selected groups  
*Can determine “causal” effects on the outcomes.  
*Generally more complex, time-consuming, and costly than other studies.  
*Mostly quantitative, but can include qualitative measures |
| Case Studies                     | *Examines usage of the product and its outcomes (e.g., usability, participant attitudes, implementation fidelity, student engagement and learning) in one or more sites (schools or classrooms)  
*Emphasizes qualitative measures but can also include quantitative measures  
*Uses rigorous and systematic methodology  
*Focuses strongly on the “cases,” but lacks a control group |
| Short-Cycle Evaluation Studies   | Similar to the Case Study, but tends to be:  
*Shorter in duration (12-14 weeks)  
*Simpler in methodology and less costly  
*More participatory with the developer and the practitioners  
*Highly practical to yield a timely assessment of a product’s potential for broader use |
**Experimental Studies**

At the top of “rigor hierarchy,” experimental designs offer the advantage of comparing the ed-tech product to a “control” condition. The latter is normally the “regular” program or curriculum, but it could also be another recently introduced ed-tech product. The most commonly used experimental designs are:

- **Randomized controlled trials (RCTs):** Sometimes referred to as the “gold standard” for evaluations, RCTs randomly assign schools, classrooms, and/or students to the experimental and control conditions.

  *Example:* There are six fifth-grade classrooms at Clover Elementary School. Three are randomly assigned to use the ed-tech product of interest, “Science Maker,” and the other four to use the existing laboratory curriculum.

- **Quasi-experimental designs (QEDs):** At a less rigorous and more practical level, QEDs compare the experimental group to a control group that is similar in characteristics but not randomly composed.

  *Example:* There are 10 middle schools in the Metro Public School District. Four of the schools elect to use the ed-tech product, “Writing Wizard.” Of the six remaining schools, the four most similar to the experimental sample are selected to serve as control schools.

---

**Strengths**

Experimental studies, if well designed and conducted, offer the following advantages:

- **Control:** Because a similar comparison group (not using the product) is employed, there is control over contaminating or biasing factors such as student ability, student demographics, school characteristics, etc.

- **Rigor:** Positive outcomes can be attributed to the product with more confidence than for other designs

- **Credibility:** Experimental studies are highly regarded by researchers and policy makers.

---

**Weaknesses**

- **Scarcity:** Experimental studies are relatively costly and time-consuming to conduct. In a rapidly growing ed-tech industry, relatively few products will have evidence from such designs.
- **Applicability**: Many ed-tech instructional products are largely or exclusively used as supplements to broader school curricula, thereby limiting their classroom time or “dosage.” Many other ed-tech products are non-instructional, used by teachers for data management, assessments, and authoring lessons. Finding differences on achievement tests or other educational outcomes therefore would be quite challenging and seemingly unreasonable to expect.

- **Timeliness**: Compared to other designs, experimental studies take more time to complete and disseminate publicly. Those appearing as journal articles (a positive attribute for credibility!) typically will take even a few years longer to appear on the scene. So, consumers of evidence need to consider whether the findings presented are relevant to the educational conditions currently existing at their schools, such as the adoption of Common Core State Standards, increased access to technology and bandwidth, hiring of teachers and principals trained in new types of preparation programs, etc.

- **Realism**: Experimental studies (particularly RCTs) may create some degree of artificiality in collecting evidence. In real-life, school districts and individual schools become interested in ed-tech products and try them out. Much less natural is to have a coin flip designate one’s role as either a product user or control group member.

**Case Studies**

Similar to experiments, “Case Studies” also constitute a broad category. Such studies can be small or large in the number of classrooms (students) served, simple or complex in the types of measures used, or highly rigorous or more informal in methodology.

*Example*: An ed-tech literacy product, “Really Ready Readers,” provides a library of reading texts for elementary students. Based on students’ assessed skills, students access texts at varied difficulty levels to read on their individual device and respond to interspersed reading comprehension assessment items. Feedback on texts read and performance is provided to students, parents, and teachers. A case study conducted by a university research center describes the implementation of the program in fourth-grade classes in three elementary schools in the same district. Lacking a control group comparison, the study focuses on how the product was used by teachers and students, their perceptions of the product, and informal outcomes such as number of texts read at different levels, comprehension scores, etc. The conclusions strongly support the product with regard to technical support, design, ease of use, and participant reactions.

The above case study seems likely to provide useful and credible evidence on the product for several reasons:

- Conducted by an apparently neutral evaluator
- Includes systematic data collection using multiple measures with multiple classrooms
Exercises implementation processes, participant experiences, and learning outcomes associated with the product.

But beware of case studies that are weaker due to the following characteristics:

- Very small sample or timeframe (e.g., only one teacher and/or only a few class periods)
- Conducted by the developer or by an evaluator having personal stake in the results
- Poorly described or overly informal (not rigorous) data collection methods
- Participants that are typical of regular teacher users, e.g., highly skilled in technology or connected in some way to the developer or product

Strengths

Case studies, if well designed and conducted, offer the following advantages:

- **Realism ("external validity")**: A meaningful picture is provided of how the ed-tech product is implemented and its impacts on participants regarding engagement, attitude, and learning outcomes.
- **Rigor**: Although a control group is lacking, rigorous data collection methods increase confidence that the results are accurate.
- **Comprehensive**: A strong case study includes multiple (mixed) quantitative and qualitative (user perception) measures of implementation processes and outcomes.
- **Readability**: Case studies tell a meaningful story about practitioners’ experiences with the product.

Weaknesses

- **Subjectivity**: Case studies are more subjective and “qualitative” in nature than experimental studies. Biases of participants may influence results.
- **Limited sampling**: Given the extensive data collection, case studies are typically limited to a few sites (e.g., classrooms or schools) and thus may have limited generalizability.
- **One-treatment only**: There is no control group and therefore no basis for making inferences about the benefits of the product relative to existing approaches.

Short-Cycle (Pilot) Studies

“Short-Cycle Evaluation Studies” (SCES), although less common than experimental studies and case studies, are receiving increasing attention today as vendors and practitioners share interest in the discovery and evaluation of ed-tech products. Importantly, in the interests of
maintaining high evidence standards, we distinguish SCES from what educators may refer to as “tryouts,” “demos,” “showcases,” “expos,” and the like. These labels denote relatively limited, informal interactions with ed-tech products rather than systematic studies (Morrison et al. 2015). For example, several teachers might review a particular product and then, with or without fully implementing in their classes, offer impressions about its appeal to the district technology director. Another example is a developer visiting the school district and giving a demonstration of the ed-tech product to selected teachers and administrators. A third might be a “tradeshow” event where several developers set-up booths to display and describe their products.

As a more structured alternative to these examples, SCES is much more similar to a case study than to an experimental study. It focuses on capturing realistic efforts by practitioners to use an ed-tech product to improve or enhance classroom instruction. Accordingly, there will not be a control group or a goal of proving product effectiveness in raising student achievement. Emphasis will be placed most directly on user impressions and experiences regarding product usability, quality, technical support, and benefits for teaching and learning. Multiple data sources such as surveys, observations, interviews, usage activities, and student performance are also included as much as feasible. School districts may not have the resources or expertise to conduct such studies on their own. Help from an independent evaluator can be advantageous in producing a stronger, more credible study.

Where SCES differs from case studies is in the following features:

- **Quick**: “Short cycle” means exactly that. While case studies may last an entire year or even longer, SCES generally run about a semester at most—10 to 14 weeks.
- **Economical**: The SCES concept is to provide evidence of the “potential” efficacy of multiple products, as needs arise in a school district. Costs will need to be paid by providers, the school district, or shared between them. But neither is likely to have much room in their budget to fund such studies (especially, smaller districts and smaller ed-tech companies). Thus, costs must be kept low relative to case studies and experimental studies.
- **Matching Needs and Products**: The SCES examines ed-tech product being used to support defined instructional needs. To help ensure more meaningful reviews, the instructional needs of participating teachers and product applications must be closely matched. Simply put, teachers wanting help in mathematics should be matched to products that address that need, not ones in other subjects or other application domains (e.g., data management).
- **“Teacher Evaluators”**: Implementation of SCES requires a district to engage teachers who are willing and able to spend time evaluating products. While a case study likely will include all or most teachers using the product at the target sites, participation in SCES occurs as a non-required, supplementary activity for
the SCES time period. Thus, SCES participants will usually and desirably be teachers having special interests in using and exploring available ed-tech products.

- **Participatory Approach:** Although a primary purpose of SCES is to judge the ed-tech product’s quality and utility, a related function is formative evaluation and product development. Even the best designed product may lack features needed locally and more broadly to be maximally effective. Accordingly, the SCES not only involves product users (teachers and students) but also the developers and evaluators in communicating findings, impressions, and improvement needs both during and following study completion.

**Example.** Recently, we have piloted SCES methodology in a project for the NYC Board of Education’s Innovation Zone (“i-Zone”) initiative. The approach used was favorably received by both teachers and vendors. Main components included:

- Soliciting participation by vendors interested in having their products piloted by practitioners and evaluated for potential for wider district use.
- Selecting products that both (a) address high-priority district needs for instructional or administrative support and (b) are at reasonably advanced levels of development with regard to design, operation, technical support, and track record.
- Matching the selected products to expressed interests by participating schools and their teacher review teams.
- Providing the teacher teams with appropriate professional development (PD) and resources for using the products over a short-cycle (12-week period).
- Having the independent evaluator:
  - Obtain informal feedback from teachers and vendors mid-way through the SCES to identify any problems inhibiting successful product usage
  - Observe teachers and students using the product in an actual class
  - Interview and survey teachers, students, and developers at the end of the cycle
  - Examine usage data (e.g., log-in, instructional time, components accessed, etc.)
  - Examine performance data (e.g., scores on embedded questions or quizzes, units completed, learning objectives mastered)
  - Prepare draft report on findings and recommendations for review by teachers and developers
  - Prepare final report after considering feedback.
Strengths

Short-cycle evaluation studies, if well designed and conducted, offer the following advantages:

- **Quick turnaround**: The studies are completed and recommendations about product quality and potential are made within 10-14 weeks.
- **Economical and Efficient**: Studies are completed with minimal burden on participants and low cost per product.
- **Developmental**: Providing formative feedback for product improvement locally and generally is highly emphasized.
- **Participatory**: Studies directly involve end-users (teachers), developers, district educational technology staff, and external evaluators in planning, execution, and interpreting results.

Weaknesses

- **Subjectivity**: Similar to case studies, SCESs are more subjective and “qualitative” in nature than experimental studies. Biases of participants may influence results.
- **Limited sampling**: Typically, a SCES for a particular product involves only a small number of teachers and schools. Compared to typical teachers, they are likely to be more motivated and prepared to employ technology in teaching.
- **Restricted timeframe**: The intentional short duration of the tryout limits outcomes to what occurs in the short run rather than over an extended implementation period.
- **Quick takeoff**: SCES requires product usage to be up and running fairly immediately in the participating classrooms without the benefit of district or whole-school implementation support (e.g., via extended professional development and peer coaching).

So How Much Should We Trust This Evidence?

A Practical Evaluation Tool

The previous sections have provided a “short course” on common approaches to evaluating ed-tech products. Each type of study was characterized by different components, strengths, and weaknesses. Here, we will assemble the pieces in a practical tool for judging the usefulness of evidence presented in evaluation reports.
# EVIDENCE EVALUATION RUBRIC

<table>
<thead>
<tr>
<th>Criterion</th>
<th>RATINGS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Low</strong></td>
<td><strong>Moderate</strong></td>
</tr>
<tr>
<td><strong>Study Rigor</strong></td>
<td>Weak or unclear evaluation design:</td>
</tr>
<tr>
<td></td>
<td>• Poorly described methodology</td>
</tr>
<tr>
<td></td>
<td>• Lack of similar control group in experimental study</td>
</tr>
<tr>
<td></td>
<td>• Limited, unsystematic data collection in case study or short-cycle evaluation study</td>
</tr>
<tr>
<td></td>
<td>• Outcome measures without clear validity or biased in content toward ed-tech product’s contents</td>
</tr>
<tr>
<td></td>
<td>• Inadequate or biased analysis or interpretation of results</td>
</tr>
<tr>
<td><strong>Evaluator</strong></td>
<td>In-house evaluator or external evaluator who appears to have self-interest in the product.</td>
</tr>
<tr>
<td><strong>Recency</strong></td>
<td>Evidence is dated (over 5 years old)</td>
</tr>
<tr>
<td>** Authentic**</td>
<td>Study conditions are dissimilar to those likely to occur in the school district:</td>
</tr>
<tr>
<td></td>
<td>• The teachers (e.g., specially trained or expert.)</td>
</tr>
<tr>
<td></td>
<td>• Types of students (e.g., special classes)</td>
</tr>
<tr>
<td></td>
<td>• Structure (e.g., schools or teachers select and use the product under special incentives).</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
• Supports (e.g., unusual amounts of PD and/or other resources are provided)

<table>
<thead>
<tr>
<th>Meaningful Outcomes</th>
<th>The outcomes measured are weak or limited in meaning due to one of more of the following:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Giving unfair advantage to the particular product (e.g., assessing only learning specific to program content)</td>
</tr>
<tr>
<td></td>
<td>• Being restricted to very short-term achievement (pretest-posttest) gains</td>
</tr>
<tr>
<td></td>
<td>• Being restricted to brief, isolated measures (a short quiz or teacher survey).</td>
</tr>
<tr>
<td>Results address reasonable educational outcomes for the scope of the study, including data on implementation and user reactions.</td>
<td></td>
</tr>
<tr>
<td>Results comprehensively address ease of implementation, participants’ satisfaction, and reasonable educational outcomes that fit the duration and intensity of product use.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Report Source</th>
<th>Disseminated by provider (not available from external agency)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Disseminated by university, research organization, federal government, school district, or other credible external agency.</td>
</tr>
<tr>
<td></td>
<td>Peer-reviewed journal publication</td>
</tr>
</tbody>
</table>

**Making the Evidence Judgment: Should We Use it or Lose it?**

Evidence may be reviewed in various forms—a full technical report or journal article, a research brief, a PowerPoint presentation, or some other abbreviated form. Rarely would you need to read a complete report to obtain the information needed for applying the above evaluation criteria. If there is an Executive Summary (usually 3-5 pages long), it will likely suffice. Should additional information be needed, you can probably find it fairly quickly in the “methodology” section of the main report.

Once you’ve obtained a reasonable picture of the study, it shouldn’t take long to complete the Evidence Evaluation Rubric Ratings. Looking back at the tool, we can picture rare cases where every rating is the same—high, medium, or low. But more typically, we will end up with an assortment of rating levels. What to do then? Take an average? Assign points and
cut-off scores for the three rating levels? We recommend more of a “decision-tree” approach given that not all criteria have equal importance. For example, evidence from a study having low rigor should be rated low regardless of other qualities—simple as that! Less simple is deriving usable formulae to yield a decision for every possible combination of ratings. However, you should find the following guidelines fairly quick and easy to employ:

**Guideline 1:**

<table>
<thead>
<tr>
<th>Study Rigor</th>
<th>Low</th>
<th>Moderate</th>
<th>High</th>
<th>Overall Evidence Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td>LOW</td>
</tr>
</tbody>
</table>

If Study Rigor is *LOW*, overall evidence quality should be rated *LOW*, regardless of all other ratings. This rule may sound rather harsh, but think about it—if the study that produced the evidence is clearly suspect to begin with, the other attributes (e.g. Evaluator, Recency, Source, etc.) shouldn’t matter. For example, depending on the particular journal, even a published study can be low in rigor.

**Guideline 2:**

<table>
<thead>
<tr>
<th>Study Rigor</th>
<th>Low</th>
<th>Moderate</th>
<th>High</th>
<th>Overall Evidence Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td>LOW</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2+ Other Criteria</th>
<th>Low</th>
<th>Moderate</th>
<th>High</th>
<th>Overall Evidence Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

If Study Rigor is * Moderately*, and two or more other criteria are rated as *LOW*, an overall rating of *LOW* should be considered.
Guideline 3:

<table>
<thead>
<tr>
<th></th>
<th>Low</th>
<th>Moderate</th>
<th>High</th>
<th>Overall Evidence Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study Rigor</td>
<td></td>
<td>[✓]</td>
<td></td>
<td>MODERATE</td>
</tr>
<tr>
<td>Most Other Criteria</td>
<td></td>
<td>(OR) [✓]</td>
<td>[✓]</td>
<td></td>
</tr>
</tbody>
</table>

If Study Rigor is MODERATE, and most other criteria are rated as MODERATE or HIGH, an overall rating of MODERATE should be considered.

Guideline 4:

<table>
<thead>
<tr>
<th></th>
<th>Low</th>
<th>Moderate</th>
<th>High</th>
<th>Overall Evidence Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study Rigor</td>
<td></td>
<td>[✓]</td>
<td></td>
<td>HIGH</td>
</tr>
<tr>
<td>All Other Criteria</td>
<td></td>
<td></td>
<td>[✓]</td>
<td></td>
</tr>
</tbody>
</table>

If Study Rigor is MODERATE, and all other criteria are rated as HIGH, an overall rating of HIGH should be considered.
Guideline 5:

<table>
<thead>
<tr>
<th></th>
<th>Low</th>
<th>Moderate</th>
<th>High</th>
<th>Overall Evidence Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study Rigor</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>All Other Criteria</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>(OR) ✓</td>
</tr>
</tbody>
</table>

If Study Rigor is HIGH, and all of the other criteria are rated MODERATE or HIGH, an overall rating of HIGH should be strongly considered (unless a particular criterion is unusually important to your district).

Every evaluation study will be unique in characteristics. The above Guidelines, therefore, are only suggested pathways for decision making. Clearly, the primary criterion is Study Rigor. Once it is classified, using the Guidelines and personal judgments should be easy and successful

__________________________________

For more information, contact Steven M. Ross (sross19@jhu.edu) or Jennifer Morrison (jrmorrison@jhu.edu), Center for Research and Reform in Education (CRRE), Johns Hopkins University. 300 E. Joppa Rd., Towson, MD 21286,