Assessing Relevance and Credibility of Available Evidence[1]

In some cases, existing studies could be useful to evaluate whether educational programs and strategies (Solution Options in DecisionMaker) that have been implemented in other contexts would work for a decision-maker's own context. However, decision-makers may struggle to determine whether a study is indeed relevant to their own context and how seriously they should take the findings given varying levels of quality or credibility. Decision-makers may want to consider results from a less rigorous study, but may want to account for the fact that that the study may over- or under-estimate the effectiveness of the program being studied.

To account for these issues, DecisionMaker provides two indices developed by CBCSE that can be used to evaluate the relevance and credibility of studies: the Relevance Index and the Credibility Index. We recommend you review the items in each index below before reading each study you want to consider as evidence so you know what to look out for.

Key terms

A **target population** is the complete collection of students/teachers/schools or other units that we want to study.[2] For example, the target population of a study on algebra skills of high students in a particular school district may be all 150,000 students in every high school in the district.

An **outcome** is a change or impact caused by the program being evaluating, or it could be a characteristic of the respondents you want to measure.[3] An outcome of a reading program for third graders in an elementary school might be to raise reading skills among third grade students.

Measures are the items in a research study to which the participants respond[4], which are used to assess performance on the outcomes of interest.

[1] The design of these indices was influenced by a number of sources: Digital Promise's "Evaluating Studies of Ed Tech Products" , NESTA's "Standards of Evidence", Alliance for Useful Evidence's "What Counts as Good Evidence", and David Gough's "Weights of Evidence: The Appraisal of the Quality and Relevance of Evidence."
[2] Lohr. Sharon L. Sampling Design and Analysis. 2nd ed. 2010.
[3] https://www.povertyactionlab.org/research-resources/measurement-and-data-collection
[4] http://www.uniteforsight.org/research-methodology/module4

A. Worksheet to Assess Relevance of Available Evidence to Your Purpose and Context:

First, establish if prior evidence exists:

1. Are there prior studies conducted on this educational program or strategy (Solution Option)?
    a. If yes, do those studies provide information on evidence of effectiveness or fidelity of implementation?

        i. Evidence of effectiveness Yes No

        ii. Feasibility of implementation Yes No

    b. If no, you may need to plan to collect your own data.

If you find studies that provide information on evidence of effectiveness or feasibility of implementation, first use the Relevance Index below to establish whether that information is relevant to your context. Afterwards you can use the Credibility Index to evaluate any studies you deem relevant to your context.

2. To help you determine whether there is at least one study that was conducted in a similar context to your school/ district/ state:

1. Choose the most important factors to consider for your decision problem from the **Relevance Index** below by filling the relevant radio buttons
2. Read each study carefully and score each one 0-3 for similarity of study context to your own on each factor you selected as important for you *(3= very similar, 2 = moderately similar, 1 = slightly similar, 0 = not at all similar)*
3. Divide the total score earned by each study by the total maximum possible score on this index and then multiply by 100 to give you a Relevance Index between 0 and 100.
4. Compare the overall score for each study with the interpretation table below to come to a conclusion about which, if any, of the studies are relevant enough to your context.
5. Record your assessments of each study in the **Relevance Summary Table** below.

Relevance Index

Contextual Factor	***Things to look for and think about***	***Important consideration for me as to whether this study is relevant to my context***	***Study context is similar to mine*** *(3= very, 2 = moderately 1 = slightly, 0 = not at all)*
Recency of study	Extent to which the Solution Option is still likely to be applicable in my context today	o	
Student demographics	Age of students on which the program/ intervention was tested	o	
	Baseline performance of students before implementing the program	o	
	Percent economically disadvantaged, e.g., as indicated by FRPL	o	
	Percent classified as minority	o	
	Percent ELL	o	
	Percent Special Education	o	
School Context	Charter vs. district school vs. private	o	
	Selective vs. open to all applicants	o	
	Characteristics of teachers (degree level, certification, tenure status, average experience)	o	
	Availability of necessary technology	o	
Relevance of measure used	Whether the outcome measure reported in the study is relevant for your goals, e.g., if you are trying to reduce behavior incidents, you might find a study that reports suspension rates to be relevant, but a study that reports attendance would not be as relevant	o	
		Total possible score: Y= # of factors selected * 3	**Total score for this study: (X)**
Relevance Index			**(X/Y) x 100**

Note. FRPL = Free and Reduced Price Lunch; ELL = English Language Learner.

Relevance Index Interpretation Table

Relevance Index (= Total Score for this study/Total Possible Score x 100)	**Relevance Rating**
Less than 30%	Low Relevance
31 - 69%	Moderate Relevance
70% or higher	High Relevance

Relevance Summary Table

Name of study	**Authors**	**Year of study**	(Column D) **Relevance Score**	(Column E) **Total possible score**	**Relevance Index** (Column D/ Column E)	**Relevance Rating** (High, Moderate, Low)	**Use this study as relevant evidence for this decision** (Yes/No)

Next Steps

We recommend you move studies that earn a rating of High or Moderate relevance forward to determine whether they are credible. If you do decide to move forward with a study of low apparent relevance to your context, you may want to be sure that it scores high on credibility in the next step.

If none of the studies you review score better then Low Relevance, you may want to consider designing your own study and collecting your own data.

The output of Step A on Relevance is to provide a Relevance rating for each study you review, and a decision on whether to use any of these studies as evidence for the Solution Option(s) you are considering, or to collect your own data. If one or more studies pass the Relevance threshold (greater than 30%), then you can proceed with the relevant study to the next step (on credibility). If no study passes the threshold, then our recommendation would be to collect your own data on that Solution Option. Our Relevance threshold is merely a guideline, but you can choose your own threshold by which to accept studies that makes sense given the context and Decision Problem at hand.

B. Worksheet to Assess the Credibility of Available Evidence

If you have selected an Evaluation Criterion related to evidence of effectiveness, e.g., "Impact on …", then you can use the Credibility Index below to assess how seriously you should take the findings of each study you reviewed which passed the 'Relevance threshold' in Part A.

Credibility Index

Part I: Source of Study and Outcomes Measured

1. For each section (1-4) in Part I of the Credibility Index, select the statement that most accurately reflects the study in question. Select ONE statement only per section in Part I.
2. Add up the points you awarded in the last column for all four sections.

Part I	***What to look for and think about***	***Select one statement per section below that most accurately reflects the study in question***	***Possible points***	***Add possible points for each selected statement***
Section 1: Who conducted the study?	It is not clear who conducted the study.	o	-1	
	The study was conducted by the program vendor.	o	0	
	The study was conducted by an external evaluator hired by the program vendor.	o	1	
	The study was conducted by an external evaluator acting as an independent third party, i.e., not paid by the vendor. This may include collaborations between the evaluator and implementing partners such as school districts or other educational or research institutions, but excluding the program vendor.	o	2	
Section 2: Who published the study?	It is not clear where the study was published.	o	-1	
	The study was published by a vendor.	o	0	
	The study was published by a third party (i.e., other than the vendor) but not a peer-reviewed journal. E.g., a university, research organization, government, school district. Keep in mind some technical reports are also later published in a peer-reviewed journal, so you may wish to check to see if there is also a published version.	o	1	
	The study was published in a peer-reviewed journal.	o	2	

Section 3: Length of exposure to the program	Length of exposure is not clear from the study.	o	-1	
	Length of exposure is too short to make a difference.	o	0	
	Length of exposure is too long to reflect likely effect in regular practice.	o	1	
	Length of exposure is about right.	o	2	
Section 4: Meaningful outcomes	It is not clear which outcomes are being measured, e.g., it is not clear whether the study is evaluating geometry skills or algebra skills.	o	-1	
	The outcomes measured are not at all aligned with the ultimate goal for implementing the intervention, e.g., the study investigates whether an after-school supplemental math program improves geometry skills, despite the fact that the program aims to improve algebra skills.	o	0	
	The outcomes measured only capture short-term behavioral changes, but not the longer-term educational outcomes that you are interested in. E.g., the study only documents whether students are attending an after-school math program, but does not measure whether their math skills are improving.	o	1	
	The outcomes measured are aligned with some but not all of your ultimate goals for implementing the intervention, e.g., the study is measuring algebra skills when the primary goal of the program is to improve both algebra and geometry skills.	o	2	
	The outcomes measured are aligned with all of your ultimate goals for implementing the intervention, e.g., the study is measuring algebra skills when improving algebra skills is the primary goal of the program.	o	3	
Total for Part I		**Add up points from Section 1-4**	**Possible: 9**	

Part II: Study Sample

Imagine that you are the superintendent of a large, diverse school district, and you want to investigate the social and emotional competencies of all high school students in your district. However, the school district contains 25 high schools, with a large student population of 150,000 students. You have a limited budget, and decide to ask a research team to collect the data for you. It might not be feasible to go into all 25 high schools in the district and get all 150,000 students to take an assessment of social and emotional competencies, so the research team will likely select a subset of students, i.e., a sample, to make the data collection process more feasible.

Key terms

A **sample** is a subset of a population, for example, a subset of the students/teachers/schools that make up a target population for an evaluation.[5]

A sample is **representative** if the sample is similar to the target population on all important characteristics.

Sample Size: The number of units (e.g., students/teachers/schools) in a sample.

Statistical power: The probability that the estimate of the program effect will be found statistically significant if an effect of that size is determined to have occurred.[6]

There are two concerns with drawing a sample in order to get trustworthy results in an effectiveness study:

- *Is the sample representative?*
  - *For example, if the district is 50% FRPL, 75% minority and 13% ELL, the researchers should aim to draw a sample that has a similar distribution of these characteristics.*
- *Is the sample large enough to detect an effect when indeed there is one?*
  - *For example, if the sample only has 10 participants and the study aims to measure outcomes for social emotional competencies, this would probably be too small a sample. However, if the study aims to measure ease of implementation across 10 different classrooms, this sample size would be more reasonable.*

[5] Rossi, Lipsey & Henri. Evaluation: A Systematic Approach. Eight edition. 2019.
[6] Rossi, Lipsey & Henri. Evaluation: A Systematic Approach. Eight edition. 2019.

3. Assess the study in question on its representativeness and sample size in Part II of the Credibility Index below. Select ONE statement only per section in Part II.
4. Add up the points you awarded in the last column for the two sections.

Part II:	***Things to look for and think about***	***Select one statement in the relevant section below***	***Possible points***	***Add possible points for each selected statement***
Section 5: Representativeness: How much does the study sample mirror your target population? Think about the key factors that matter for what the study is trying to measure – does the sample have a similar distribution of these factors compared with the target population?	The characteristics of the study sample are not clear.	o	-1	
	The study sample is not representative of my target population.	o	0	
	The study sample is moderately representative of my target population.	o	1	
	The study sample is highly representative of my target population.	o	2	
Section 6: Sample size Does the size of the sample, i.e., the number of participants in the study, seem reasonable? (For researchers: do you think there is enough power to detect an effect if indeed there is one?	The size of the study sample is not clear.	o	-1	
	The study did not have a reasonable number of participants.	o	0	
	The study had a fairly reasonable number of participants.	o	1	
	The study had a very reasonable number of participants.	o	2	
Total for Part II:			**Possible:** 4	

Part III: Rigor of Methodology

Key terms

A comparison group is a group that did not receive a program and can be used to compare against the group that did receive the program. The key challenge in a good efficacy study is to find a group that did not receive the program, but closely resembles the group that did receive the program, meaning both groups should be similar on average across all main observable characteristics, such as student demographics or school characteristics.

To isolate the effects of a social program, researchers conducting effectiveness studies need to measure the outcomes for the individuals exposed to the program (the "treatment group") and find a credible way to estimate the outcomes that would have occurred in the absence of the program.[7] *To do so, researchers must identify a comparison or "control" group that is similar to the group exposed to the program except for participation in the program.*

There are two considerations:

- *The most important consideration is identification of a credible comparison group. A credible comparison group is one that is similar to the group that received the program on characteristics that are relevant for the goals of program. To assess how credible a comparison group is, ask **how** program participants were selected to participate in the program:*
    - *Were program participants selected because of certain student, teacher or school characteristics? For example, were students lagging behind in literacy chosen to participate in a reading program? If so, does the comparison group perform at the same baseline reading levels as program participants? Or were schools with motivated principals and strong infrastructure selected to participate in the program? If so, does the comparison group have equally motivated principals and similar infrastructure to the program schools?*
- *The second consideration is whether data on outcomes are collected multiple times (e.g., before and after the program, and on subsequent occasions) for the treatment and the comparison group. This can be useful if you want to account for baseline differences between the two groups, or if you want to measure longer term outcomes.*

[7] Rossi, Lipsey & Henri. Evaluation: A Systematic Approach. 8th edition. 2019.

5. Select ALL statements in Part III of the Credibility Index below that apply to the study in question.
6. Add up the points you awarded in the last column for Section 7.

Part III:	***Things to look for and think about***	***Select all that apply*** *(Radio button)*	***Possible points***	***Add possible points for each selected statement***
Section 7: Rigor of methodology for evaluating outcomes	1. Determining whether there is a credible comparison group			
	1a. First determine whether there is a comparison group of any kind.			
	It is not clear whether there is a comparison group.	o	-1	
	There is no comparison group.	o	-1	
	The study includes a comparison group which does not participate in the program being studied.	o	1	
	1b. Then, identify how that comparison group was selected.			
	It is unclear how program participants were selected.	o	-1	
	Program participants were selected based on certain observable characteristics (e.g., gender, academic performance), and the comparison group is not similar on those characteristics. For example, the lowest-performing students were selected to participate in a reading program, and the comparison group includes high-performing students.	o	0	
	The study compares outcomes for students/teachers/schools who are receiving the program with outcomes for counterparts who have similar characteristics but are not participating in the program. It may do so either by identifying a comparison group that shares several known characteristics with the program participants, e.g., same grade, gender, SES (statistical matching), or by first matching program participants with non-participants who could have been just as likely to participate in the program, as predicted by known characteristics such as age and gender, and then comparing outcomes for the matched pairs (propensity score matching techniques).[8]	o	1	

[8] https://www.povertyactionlab.org/sites/default/files/resources/2016.08.31-Impact-Evaluation-Methods.pdf

	The intervention is provided to students/teachers/schools who are above a cutoff point for eligibility. The study compares participants who are just above the cutoff, and therefore receive the intervention, with students/teachers/schools who are just below the cutoff, and therefore do not receive the intervention. This design should ensure the two groups are highly comparable.	o	3	
	The study uses a randomized controlled trial (RCT) in which students/teachers/schools are chosen at random to either participate in the program or to serve in a comparison group.	o	5	
	2. Measuring outcomes over time			
	The study includes before and after measures, e.g., a pre-test/survey/observation before the intervention and a post-test/survey/observation after the intervention.	o	1	
	The study includes a second post-test several months after the intervention ends.	o	1	
	The study assesses outcomes multiple times before/during and after the intervention.	o	1	
Total for Part III		**Add points for Part III**	**Possible: 10**	

7. Now add up the points earned by this study for Parts I, II and III of the Credibility Index
8. Use the Credibility Index Interpretation Table below to find the credibility band your score falls into.
9. For low or moderate credibility studies, you can multiply the effect size found in the study you reviewed by the Credibility Parameter to adjust it downwards.

Summary Score for Credibility Index

Total for Parts I-III:		
Total for Part I (possible 9)		
Total for Part II (possible 4)		
Total for Part III (possible 10)		
TOTAL	**Possible:** 23	

Credibility Index Interpretation Table

Credibility Index (Total Points for Parts I, II, III)	**Credibility Rating**	**Credibility Parameter**
Less than 8	Low credibility	0.2
8-14	Moderate credibility	0.6
15-23	High credibility	1

10. Use the Relevance and Credibility Summary Table below to document your assessments of each study you reviewed.

Relevance and Credibility Summary Table

Name of study	**Authors**	**Year of study**	**Relevance Index**	**Relevance Rating** (High, Moderate, Low)	**Use this study as relevant evidence for this decision** (Yes/No)	**Credibility Index** (Total Score adding Parts I, II, III)	**Credibility Rating** (High, Moderate, Low)

The output of Step B on Credibility is to assign a low, medium or high parameter for the credibility of a study assessing evidence of effectiveness. This parameter will function as a weight between 0 and 1 that is multiplied by the value entered on evidence of effectiveness to weight the effect size by its credibility. Decision-makers may also consider not using studies with very low credibility.

For example, if you are trying to assess a computer-assisted learning program on the Evaluation Criterion “Impact on standardized test scores” and a study you reviewed of “Option 1” reported an effect size of 0.2 but received a Credibility rating of “Moderately credible,” this is how you would proceed:

Credibility Parameter: Moderately credible = 0.6
Impact on standardized test scores (as taken from the evaluation study) = 0.2

Multiply the effect size reported in the study by the Credibility Parameter and use the new effect size as the expected effectiveness in your evaluation table in *DecisionMaker*.

[0.2] x [0.6] = [0.12]