Considerations regarding evidence review standards when using *RCT-YES*

February 2017

Lauren Scher

Concentric Research & Evaluation

Russell Cole

Mathematica Policy Research

Work on this paper was supported through Analytic Technical Assistance & Development (ATA&D) Contract ED-IES-12-C-0057, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Decision Information Resources, Inc. (DIR) was the ATA&D prime contractor responsible for organizing the work of all contributions leading to the development of this guide.

Contents

Figures

Tables

Boxes

RCT-YES—a tool for presenting analyses aligned with evidence review standards

RCT-YES is a free software tool[1] that researchers can use, in combination with standard statistical packages, to present impact findings for evaluations of interventions, programs, and policies with treatment and control (or comparison) groups. *RCT-YES* provides automated output in the form of descriptive tables that can help researchers (1) assess whether their evaluations meet evidence standards and (2) implement any necessary adjustments to help their evaluations meet evidence standards. This output is relevant to standards across a variety of established systematic evidence reviews (see Box 1), such as the What Works Clearinghouse (WWC)[2] for interventions related to education.

Box 1. What is a systematic evidence review?

The goal of a systematic evidence review is to describe the effectiveness of an intervention of interest based on the most credible evidence available. This goal is accomplished by applying standardized rules to assess the quality of evidence from effectiveness studies of the intervention. The subset of evidence that is of sufficiently high quality is then used to describe the effectiveness of the intervention.

RCT-YES relies on design-based methods[3] to estimate average treatment effects for randomized controlled trials (RCTs) and quasi-experimental designs (QEDs). These methods can be used to estimate effects for individual-level and clustered designs; they use estimators that align with the building blocks of experiments, making them a viable alternative to traditional model-based methods, such as hierarchical linear modeling (HLM). To run *RCT-YES*, researchers can use the standard statistical packages, R and STATA.[4] *RCT-YES* is flexible and offers the following types of design and analytic options:

- Adjustments for clustering when the unit of analysis (for example, students) is not the same as the unit of treatment assignment (for example, schools or classrooms)
- Covariate adjustments
- Baseline equivalence analyses

[1] Home page of *RCT-YES* software tool.

[2] Home page of What Works Clearinghouse (WWC).

[3] For details about *RCT-YES* reliance on design-based methods, see Statistical theory for the RCT-YES software: Design-based causal inference for RCTs, Second Edition (NCEE 2015–4011).

[4] For an overview about downloading *RCT-YES*, see the *RCT-YES* download page. For specific information about R, see the Getting Started information for the R Project for Statistical Computing. For specific information about STATA, see the STATA home page.

- Multiple comparisons adjustments
- The use of weights

RCT-YES can also measure impacts for specific subgroups and compare impacts across subgroups. The strength of *RCT-YES* is in its flexibility, ease of use, ability to report impact estimates under a variety of assumptions, and presentation of automated output that aligns with the information requirements for systematic evidence reviews.

This brief provides suggestions for how *RCT-YES* users can analyze data and report findings in ways that are likely to meet or exceed the standards that define credible evidence of program effectiveness for systematic evidence reviews (see more details on these standards in the next section). The brief is intended for researchers who have experience in conducting and analyzing results from impact studies, have some familiarity with the evidence review standards most appropriate for their particular evaluation, and have reviewed the *RCT-YES* background materials.[5] To provide context, this brief first discusses common evidence standards and includes implications for how to structure and analyze a data file that would be used in coordination with *RCT-YES*. It then provides suggestions for how to use information from *RCT-YES* to report study findings and supporting documentation that are needed to determine whether the analyses meet evidence standards.

When using *RCT-YES* to present findings that are aligned with standards for high-quality research studies, researchers should consider certain issues. To demonstrate those issues, this document uses a hypothetical example of a study that measures the impact of a dropout prevention program on student test scores (see Box 2). This brief also includes screenshots of *RCT-YES* output tables to illustrate this example.

Box 2. Example: A study of dropout prevention

A researcher has implemented a study in which students in grades 9–12 were randomly assigned either to participate in a multiyear dropout prevention program provided as a supplement to their regular academic courses (the treatment group) or to participate in the regular high school curriculum (the control group). In this example, 199 students were assigned to the treatment group and 201 to the control group. The key outcomes were reading and math achievement and dropout and graduation rates over a four-year period. At the end of the second year, the researcher obtained standardized test scores for math outcomes and was interested in using *RCT-YES* to conduct an analysis that measures interim effects of the program on math achievement.

[5] See the RCT-YES support page.

Understanding evidence review standards

Over the past decade, federal agencies have placed a high value on supporting evidence-based programs and on prioritizing the support of research and the implementation of programs that are backed by strong evidence of effectiveness (see Box 3). For example, researchers seeking grant funding from the U.S. Department of Education to scale up and evaluate specific educational interventions have had to document prior evidence of intervention effectiveness based on research that meets WWC standards. Other agencies, such as the U.S. Department of Health and Human Services and the U.S. Department of Labor, have developed their own evidence standards and engaged in similar review efforts to identify and support research and implementation of evidence-based programs.

Box 3. Examples of federally supported systematic evidence reviews with standards for assessing the rigor of effectiveness studies

- What Works Clearinghouse (Institute of Education Sciences, U.S. Department of Education)
- Home Visiting Evidence of Effectiveness (U.S. Department of Health & Human Services, Administration for Children and Families)
- Teen Pregnancy Prevention Evidence Review (U.S. Department of Health & Human Services, Office of the Assistant Secretary for Planning and Evaluation)
- **Clearinghouse for Labor Evaluation and Research (U.S. Department of Labor)**

In general, review standards specify three categories of evidence quality, which, for simplicity, we characterize as high-, moderate-, and low-quality evidence.[6] To determine which rating a study may be assigned, evidence review standards generally focus on three main study features: study design, sample attrition, and baseline equivalence (Figure 1).

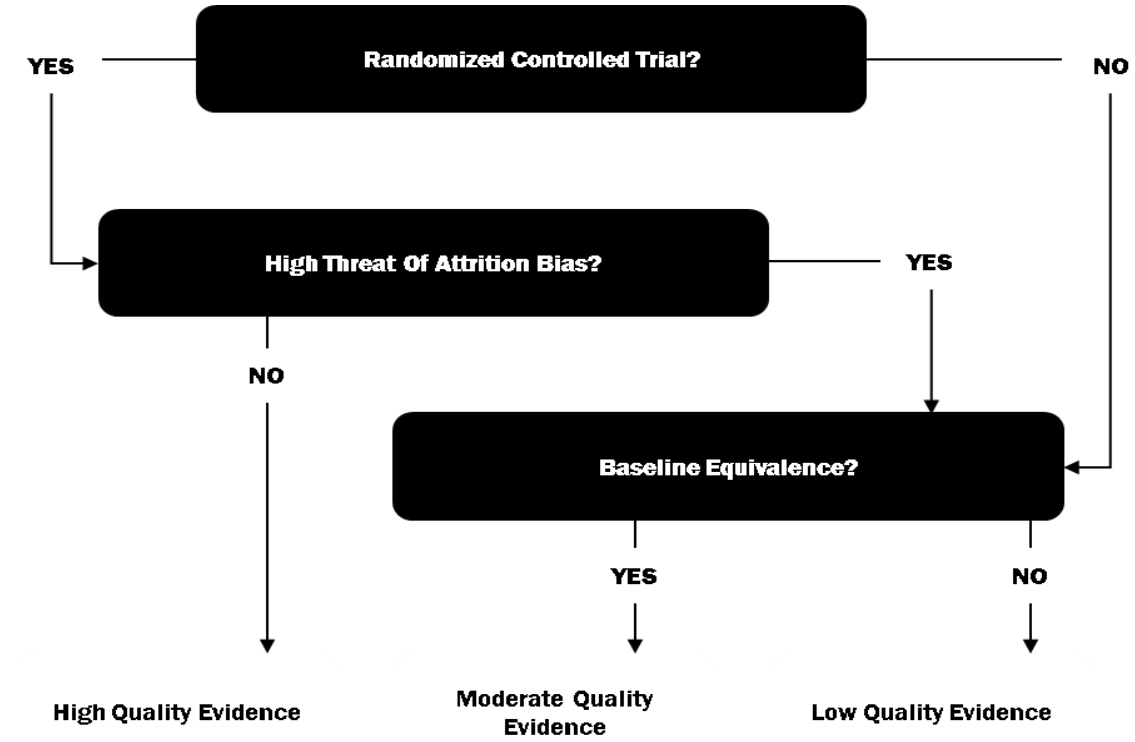
Study design. In general, to be eligible for the highest-level evidence rating, a study must use random assignment to allocate participants to treatment or control groups (that is, use an RCT design).

Sample attrition. Most evidence review standards also require that to be eligible for the highest-level evidence rating, studies must—in addition to using an RCT design—demonstrate minimal sample loss. Attrition standards vary by topic studied and are specified in advance of a review. For example, in WWC reviews, studies that examine the effects of dropout prevention programs have stricter criteria for sample attrition than studies that examine the effects of reading curricula. That

[6] Although many review standards categorize evidence into three levels of quality, terminology varies. For example, the WWC rates a study as meeting group design standards "without reservations" (the highest rating), "with reservations" (a moderate rating), or "does not meet standards" (the lowest rating). Other evidence review standards use different terms for similar ratings, such as "high," "moderate," or "low." See Teen Pregnancy Prevention Resource Center: Evidence-Based Programs.

is, given a reading curriculum study and a dropout prevention study with identical attrition rates, it is possible that the reading curriculum study might be characterized as "low attrition" but that the dropout prevention study might be considered "high attrition" because of the stricter attrition criteria for reviews of dropout prevention studies. See the Attrition Standards Brief for an explanation of how the WWC applies this standard.[7]

Figure 1. Evidence review framework



Note: This figure does not include all issues considered in evidence review standards—for example, the roles of confounding factors and outcome eligibility are not shown here. This figure highlights the issues most relevant for using *RCT-YES* to conduct analyses that meet evidence standards.

Baseline equivalence. Most systematic evidence reviews require that RCTs with high sample attrition and all QEDs use pre-intervention measures to demonstrate that the treatment and control groups are similar prior to the intervention (that is, "have baseline equivalence") in order for evidence to be considered credible. However, such evidence is deemed to be less credible than an RCT with low levels of sample attrition.[8] Studies that do not demonstrate convincing baseline

[7] Because the reason for nonresponse can vary across topic areas and sometimes may be related to the presence or absence of the intervention, the WWC sets different criteria for allowable attrition, depending on the topic. See *WWC Standards Brief–Attrition Standard* (2015). See What Works Clearinghouse, Institute of Education Sciences, U.S. Department of Education.

[8] In some cases, evidence reviews may also require that low-attrition RCTs demonstrate baseline equivalence or control for specific baseline covariates in order to be eligible to receive the highest evidence rating.

equivalence are judged as not providing credible evidence. In addition, systematic evidence reviews may sometimes require that impact analyses accommodate or adjust for baseline differences among the treatment and comparison groups to meet evidence standards. See, for example, the Baseline Equivalence Standards Brief, which describes how the WWC applies this standard.[9]

The quality of a study's results depends on design, implementation, data collection during the evaluation, preparation of data, and estimation of program effectiveness. Because the *RCT-YES* tool facilitates estimation of impacts from the data already collected for a study (that is, the tool does not play a role in study design, implementation, or data collection), it cannot be used to ensure that a study receives the highest evidence rating. However, as described below, *RCT-YES* users can follow certain analytic guidelines that will, to the extent possible, best support a moderate- or high-evidence rating.

Although *RCT-YES* is equipped to handle complex research designs, such as clustered and blocked designs, the example presented in this brief focuses on a simple RCT in which students are randomized (see "Design 1: Non-clustered, non-blocked design" in the *RCT-YES User's Manual, page 31*).[10] In the next section, we offer suggestions for carrying out an analytic plan and highlight stumbling blocks that researchers may encounter when using *RCT-YES* to produce findings intended to meet evidence standards.

Using *RCT-YES* to prepare and carry out an analytic plan

This section presents steps for carrying out an analytic plan to meet evidence review standards:

1. Read evidence review documentation.
2. Create a comprehensive analytic dataset.
3. Assess the threat of sample attrition.
4. Examine baseline equivalence requirements.
5. Conduct the most appropriate analysis, given attrition and findings of baseline equivalence.

Step 1. Read evidence review documentation

Knowing the specific requirements of a relevant systematic evidence review will increase the likelihood that the results from an *RCT-YES* analysis meet the requirements of that review effort. Therefore, you should carefully examine the evidence review standards most relevant to the topic that you are studying.

---

[9] See *WWC Standards Brief – Baseline Equivalence* (2015).

[10] With *RCT-YES*, researchers can use a variety of models and assumptions for analyses (for example, treatment-on-the-treated analyses). This brief does not discuss how these different approaches may be considered in systematic evidence reviews. For more information about the variety of options available, consult the *RCT-YES User's Manual*.

Most systematic evidence reviews publish online documents that outline the standards that are used to assess the quality of evidence. For instance, regarding our example of a dropout prevention study, researchers should review the WWC Standards Handbook[11] and the specific WWC protocol[12] for dropout prevention. Of critical importance, this protocol provides detailed information concerning attrition and baseline equivalence standards, as well as information about the kinds of outcomes included in dropout prevention reviews and how outcomes are grouped into domains. Each of these characteristics has a bearing on the strategies for using *RCT-YES*.

Step 2. Create a comprehensive analytic dataset

Chapter 4 (pages 35–43) of the *RCT-YES User's Manual* provides detailed guidance about how to construct a data file that will be compatible with the requirements of the *RCT-YES* tool. The following suggestions should help you prepare such a file:

- ***Include all individuals assigned to treatment and control groups.*** Researchers should create a comprehensive data file that includes all enrolled sample members, regardless of whether all data are available and whether all sample members are included in the analysis. With comprehensive data, *RCT-YES* can perform an accurate calculation of sample attrition (more on this in "Step 3. Assess the threat of sample attrition"). This is a particularly important issue for RCTs because a study deemed by an evidence review as having high sample attrition would be required to demonstrate baseline equivalence to meet evidence standards. If a study uses a clustered design, then the data file should include cluster-level identifiers (such as a classroom or school identification number) for each subject. *RCT-YES* will calculate attrition at the cluster level. In addition to a comprehensive data file, researchers might also consider creating a *complete case* data file that includes only sample members with non-missing baseline and outcome data (see Box 4).

Box 4. Be aware: A complete case analysis may be more likely to meet evidence standards than an analysis that includes missing data

Some evidence review standards may have strict requirements for handling missing baseline or outcome data (see the "Understand baseline equivalence requirements" section). Therefore, it might be useful to create an alternate version of your comprehensive dataset that operationalizes a "complete case analysis." To create such a dataset in a way that will allow *RCT-YES* to report accurate attrition information among the complete case sample, you should recode the observed outcome data of all observations that have missing data for key baseline variables of interest as "missing." Recoded observations with missing baseline data but observed follow-up data would effectively be treated as if they attrited from the study. *RCT-YES* analyses conducted on this complete case dataset would be more likely to meet evidence standards than analyses of the comprehensive dataset if the evidence review standards specify strict requirements for how missing baseline data are to be handled analytically.

[11] *WWC Procedures and Standards Handbook*, version 3.0 (2014).
[12] *WWC Evidence Review Protocol for Dropout Prevention Interventions*.

- ***Include all relevant data.*** The data file should include all relevant data, including baseline information (particularly those baseline variables that are most relevant to the systematic evidence review) and all outcomes and other design features required to estimate impacts (for example, randomization strata, nonresponse weights, and other analysis weights). The design features will be incorporated by *RCT-YES* into the descriptive output and impact findings.

Step 3. Assess the threat of sample attrition

For studies using RCT designs, the level of sample attrition will determine whether the highest possible evidence rating is high or moderate. *RCT-YES* provides information on sample attrition in an HTML-formatted table (see Figure 2).[13] For cluster-level designs, *RCT-YES* provides attrition information at both the individual and cluster levels. This information about treatment and control group response rates can show whether the analysis of a particular outcome meets the attrition standard. For our example, Table 1 ("Attrition Standards for Randomized Controlled Trials") on page 6 in the WWC protocol can be used to assess whether the combination of overall and differential attrition is high or low.

Figure 2. Example of *RCT-YES* output that informs the attrition calculation

Table 2. Sample sizes and key summary statistics for study outcomes for the full sample, by research status

Outcome	Number in Sample	Number with Available Data	Number with Missing Data	Percentage with Available Data	Mean	Standard Deviation
Mathematics						
Individual-level data						
Treatment group						
posttest	199	174	25	87	66.39	8.53
Control group						
posttest	201	163	38	81	64.56	8.69

Notes: Values of "-" indicate that the variable is excluded from the analysis due to the potential for the disclosure of personally identifiable information (PII). The main reasons for exclusion of an outcome variable are small sample sizes, insufficient variation across the outcome, very rare or common events for binary outcomes, or the input name for the outcome is not found in the data file. The means and standard deviations are unweighted, and are presented for binary (0 or 1) outcomes as percentages and without decimals to help minimize the disclosure of PII.

Using the *RCT-YES* table in Figure 2, *overall attrition* is calculated by first totaling the "Number with Missing Data" in *both* the treatment and control groups and dividing that total by the total "Number in Sample" in *both* the treatment and control groups. The resulting quotient should then be multiplied by 100 to get an overall attrition percentage. *Differential attrition* (difference in

[13] The formatted tables that *RCT-YES* produces are numbered in a specific and consistent order. For example, "Table 2" of *RCT-YES* formatted tables always provides information on sample sizes. "Table 8" always presents baseline equivalence information, and "Table 9" always presents impacts.

attrition rates across conditions) is calculated as the treatment group's "Percentage with Available Data," minus the control group's "Percentage with Available Data." In the example data shown in the figure, overall attrition is 15.8 percent, and differential attrition is 6.3 percent. When these results are compared to the attrition requirements stated in the WWC protocol, it is clear that attrition is "high," and the study would not be eligible for the highest evidence rating. The study would need to demonstrate baseline equivalence in order to be eligible for a rating of "meets WWC group design standards with reservations," the WWC's version of a moderate rating. Note that in situations where impact analyses include cases with missing baseline data, *RCT-YES* may present response rate information in Table 2 that is not sufficient to meet evidence review standards without additional information or analyses (see Box 5).

Box 5. Be aware: *RCT-YES* default options include imputation of missing baseline data

In its "Table 2," *RCT-YES* reports the number of individuals with follow-up data for a given outcome; however, this count may not always accurately represent the number of individuals contributing to the impact analysis as needed to calculate sample attrition in accordance with standard requirements for evidence reviews. This inaccuracy could occur because the default option for *RCT-YES* includes individuals with available outcome data but missing baseline covariates (which are imputed for the analysis, unless over 30 percent of the sample is missing the baseline covariates). As a result, if baseline data are missing, it may be useful to conduct an additional analysis using the complete case dataset because the information presented in *RCT-YES* Table 2 for such an analysis will accurately report on the analytic sample contributing to the impact estimate.

Step 4. Examine baseline equivalence requirements

For several systematic evidence reviews, high-attrition impact results from RCTs or QEDs can receive a moderate evidence rating only if the authors can demonstrate credible baseline equivalence and adjust for any lack of such baseline equivalence. Therefore, when examining evidence review standards, researchers should consider key concepts related to baseline equivalence as defined in Table 1. Each concept is further described in relation to the dropout prevention example used throughout this brief.

Table 1. Key concepts about baseline equivalence

Key concept	Definition across common evidence standards	Specific WWC Dropout Prevention example
Variables that require equivalence	Typically, equivalence must be demonstrated on a baseline measure of the outcome of interest. Some evidence review standards require a demonstration of equivalence on demographic factors as well.	The study must demonstrate equivalence of the sample on race/ethnicity, gender, at least one measure of degree of disadvantage (such as free and reduced-price lunch status), and at least one measure of academic performance.
Barometer of baseline equivalence	Many evidence standards consider the statistical significance of the difference between groups on a baseline measure to determine whether groups are baseline equivalent or not. Other standards require assessing baseline equivalence in terms of the magnitude of the difference between groups (in effect size units).	The magnitude of the baseline difference for each baseline variable must be less than 0.25 standard deviations to meet WWC standards. If the baseline difference is between 0.05 and 0.25 standard deviations, then the impact analysis must statistically adjust for the baseline variable. If differences are less than 0.05, then no covariate adjustment is necessary.
Samples used for demonstrating equivalence	Some evidence standards require a 1:1 correspondence between the sample used to demonstrate equivalence and the sample used to estimate program impacts. Other evidence standards allow some flexibility and variability in the samples used to show equivalence and the samples used to estimate impacts.	To demonstrate baseline equivalence, the study must use the same sample used to estimate impacts.
Baseline covariate imputation	Most evidence standards require a demonstration of equivalence that uses observed baseline data. In addition, some evidence standards, such as the WWC Version 3.0 standards, require that if a study must statistically adjust for baseline characteristics, then the impact analysis cannot impute any missing baseline values. Importantly, this implies that *RCT-YES*'s default approach of imputing missing baseline covariates in impact analyses may not always meet baseline equivalence requirements.	The WWC does not allow missing baseline data to be imputed for a demonstration of equivalence or for covariates used in an impact analysis. In this case, equivalence must focus on a "complete case" dataset (described in this paper) that includes only *observed* baseline and outcome data. See Boxes 4, 5, 6, and 8.

RCT-YES users can specify the baseline characteristics that they want to compare in the optional "Baseline Equivalence Analysis" input screen (see page 60 of the *RCT-YES User's Manual*). Common baseline equivalence requirements include baseline measures of the outcome, race and ethnicity, gender, age, and socioeconomic status. Because *RCT-YES* presents baseline equivalence differences in terms of effect sizes and level of statistical significance, researchers can use the criteria that are most appropriate to the relevant evidence review standards. Note that an *RCT-YES* baseline equivalence table presents a single, overall sample size, rather than a separate sample size for each variable examined at baseline, and thus, the sample sizes presented in baseline equivalence tables in *RCT-YES* do not take into account possible differences in the amount of missing data across baseline variables (see Box 6).

Box 6. Be aware: *RCT-YES* does not report line-item sample sizes used in baseline equivalence analyses

"Table 8" of the *RCT-YES* output presents baseline equivalence of the research groups using observed baseline data (no imputation). *RCT-YES* examines the equivalence of each baseline measure separately, using only the sample members that have observed data for *both* the outcome and the specific baseline measure. However, the overall sample size presented in this table is based on the maximum sample size across all variables examined at baseline; therefore, this sample size may overstate the actual sample size for each variable examined at baseline in situations where some variables contain more missing data than other variables. For example, in Figure 3, if the authors had grade-level information for the *full* sample of students with non-missing outcome data (n = 337) but were missing "percent Hispanic" data for a subset of these students, the table would present only the sample size from the grade-level variable (n = 337), even though the baseline equivalence calculation shown for "percent Hispanic" would actually be based on a smaller sample of students. Importantly, according to some evidence review standards, such as the current WWC standards, the sample used to show the effect of the intervention must be identical to the sample contributing to the baseline equivalence demonstration. As a result, and following from Box 5 ("Be aware: *RCT-YES* default options include imputation of missing baseline data"), it may be useful to conduct an additional analysis using the complete case dataset described in the section "Create a comprehensive analytic dataset." Doing so will ensure that the results in the baseline equivalence table are based on the same sample that is used to calculate the impact estimate.

Baseline equivalence results for the dropout prevention example are displayed in Figure 3, which is a screenshot of *RCT-YES*'s formatted output Table 8.

This information shows that even though baseline differences on the pretest are less than 0.05 standard deviations, the differences in grade level, percent Hispanic, percent female, and percent free and reduced-price lunch (FRL) fall within the statistical adjustment range of greater than 0.05 standard deviations and less than or equal to 0.25 standard deviations. For this reason, the study would be required to statistically adjust for grade level, Hispanic status, gender, and FRL status in order to meet WWC group design standards with reservations (adjusting for the pretest would not be required).

Figure 3. Example of *RCT-YES* output of baseline equivalence results

Table 8. Assessing baseline equivalence of the research groups in the analysis sample

Outcome and Covariate	Treatment Group Mean	Control Group Mean	Difference	Effect Size	Standard Error of Difference	p-Value of Difference
Mathematics						
posttest: Mathematics						
grade	9.82	9.75	0.07	0.08	0.09	0.459
hispanic	64	62	2	0.05	5	0.649
female	46	49	-3	-0.06	5	0.570
frl	83	77	6	0.15	4	0.166
pretest	59.56	59.66	-0.10	-0.02	0.62	0.868
Joint test						**0.665**
Sample Size	**174**	**163**	**337**			

Notes: The analysis is conducted using the full sample where cases with missing data for the baseline covariate and outcome under investigation are excluded from the analysis. The effect size is the difference between the two research groups divided by the standard deviation of the covariate for individuals in the two research groups. The findings for binary (0 or 1) outcomes are presented as percentages.

Values of "-" indicate that the baseline covariate is excluded from the analysis because the input name for the covariate is not found in the data file or due to the potential for the disclosure of personally identifiable information (PII). The reasons for exclusion due to PII-related reasons are small sample sizes, insufficient variation across the outcome, or very rare or common events for binary outcomes.

The sample size is the maximum sample size across the baseline covariates.

* Difference is statistically significant at the 0.05 level, two-tailed test.

Step 5. Conduct the most appropriate analysis, given attrition and baseline equivalence findings

As discussed earlier, evidence review standards have varying requirements for approaches to measuring program impacts, particularly regarding issues related to sample attrition and baseline equivalence. Given information on sample attrition from *RCT-YES* Table 2 and on baseline equivalence from *RCT-YES* Table 8, researchers can plan an analysis to meet evidence standards. In general, it is good practice to statistically adjust for all variables that require equivalence in an evidence review because such an adjustment may be necessary to achieve the high evidence rating.

Also, if the study shows high levels of sample attrition, it may be necessary to conduct this statistical adjustment to achieve the moderate evidence rating. In the dropout prevention study example, the study had high levels of sample attrition, which was documented in the attrition table (Figure 2); also, several variables showed a need for statistical adjustment because of baseline differences greater than 0.05 standard deviations, as shown in the baseline equivalence table in Figure 3. An impact analysis would have to adjust for these variables as covariates to produce findings eligible for the moderate evidence rating. *RCT-YES* can be used to statistically adjust for such covariates in the impact analysis.

Finally, researchers can use outcome domains to structure an approach for estimating impacts. *RCT-YES* applies multiple comparison adjustments to account for multiple hypotheses testing

within user-specified outcome domains. Therefore, researchers can group outcomes together into common outcome domains to align with established evidence review protocols (see Box 7).

Box 7. Be aware: To ensure drawing similar conclusions regarding statistical significance following multiple comparison adjustments as in a systematic evidence review, consider categorizing outcomes into domains that align with an established review protocol

Researchers may want to consult evidence review protocols to see how the review categorizes particular outcomes into common domains. In the *RCT-YES* Outcome Details Screen (see page 55 in the *RCT-YES User's Manual*), users can input multiple outcomes within a domain. *RCT-YES* will then automatically make statistical adjustments for multiple comparisons on all outcomes within the same domain. If a systematic evidence review uses a different domain categorization scheme for outcomes than the one users specify in *RCT-YES*, it is possible that the systematic evidence review will come to different conclusions about statistical significance of findings because of alternate approaches used to inform the multiple comparison adjustment.

Describing findings in a report or article

The tables and figures produced by RCT-YES can be presented as is in a report or article of study findings. The formatted output in RCT-YES's Table 9 (see Figure 4) presents impact estimates for the full sample and for any optionally specified subgroups.

Figure 4. Example of *RCT-YES* output of impact findings

Table 9. Impacts on study outcomes for the full sample and baseline subgroups

Outcome and Subgroup	Treatment Group Mean	Control Group Mean	Difference (Impact Estimate)	Effect Size	Standard Error of Difference	p-Value of Difference
Mathematics						
posttest: Mathematics						
Full Sample	66.28	64.56	1.72	0.20	0.94	0.068
Sample Size	**174**	**163**	**337**			

Notes: The impact estimates are calculated using simple differences-in-means methods or regression models that control for baseline covariates if specified. Cases with missing data for the outcome and subgroup under investigation are excluded from the analysis. The Control group means are sample means, and the Treatment group means are calculated by summing the Control group means and the impact estimates. The effect size is the difference between the two research groups divided by the standard deviation of the outcome for individuals in the Control group. All estimates are obtained using weights if specified.

The sample size is the maximum sample size across the outcomes for the full sample analysis.

Regression R^2 values for the full sample analysis are: 0.03 for posttest.

* Difference is statistically significant at the 0.05 level, two-tailed test.

^ Difference remains statistically significant at the 0.05 level, two-tailed test after applying the Benjamini-Hochberg correction for multiple hypothesis testing across all full sample analyses in the same domain.

a Indicates p-values to test for differences in impacts across subgroup categories.

For our dropout prevention study example, the impact estimate is 0.20 effect size units, which is not statistically significant at the 5 percent level. See Box 8 for further considerations in how to report findings for systematic evidence reviews.

Box 8. Be aware: Supporting your publication with complete case analyses may help it meet evidence standards

When space allows and the main analysis is not focused on the complete case dataset, we recommend presenting the complete case analysis in an appendix (or an online appendix). Doing so provides an evidence review with exactly the information that reviewers need if they determine that the main analysis does not meet certain requirements of the evidence review. That said, some journals might not provide an opportunity to include these findings, even as an online appendix. In such a case, authors should mention in the body of the article that a complete case analysis was conducted as an assessment of sensitivity or robustness and that the results are available upon request. This signals to an evidence review body that these results were estimated, which is occasionally a requirement before an evidence review has permission to query an author for the results. Regardless of whether the complete case analyses are included in the publication or not, it is important to describe to the reader the substance of the complete case findings, relative to the benchmark results. In particular, we recommend describing the direction and significance of this sensitivity analysis in the results section to indicate to your reader whether the findings are substantively similar across benchmark and complete case specifications.

Conclusion

In summary, this brief highlights the nuances of using the *RCT-YES* software package to analyze and report impact findings that can potentially meet evidence standards. *RCT-YES* presents average treatment effects of interventions, programs, and policies by using methods vetted by a panel of experts in the field of causal inference. *RCT-YES* has flexible options that can be invoked to satisfy reporting requirements for systematic evidence reviews. The default approaches for estimating program impacts in *RCT-YES* do not always ensure that the analyses and results will meet evidence standards. *RCT-YES* cannot "fix" a study that was not designed or implemented well. Whether a study meets evidence standards will depend on multiple features of the study, including design, implementation, data collection, and analysis.

Assuming that researchers have used appropriate study design, implementation, and data collection procedures, they can also take specific steps to increase the likelihood of conducting analyses and generating results from *RCT-YES* that will meet standards for high-quality research studies. In particular, *RCT-YES* users should understand the evidence standards most relevant to the topic that they are studying, make sure that study attrition can be assessed (for example, by including the full sample in the *RCT-YES* dataset), conduct appropriate baseline equivalence analyses, and adjust for baseline covariates as necessary. In addition, users should also consider estimating impacts by using complete case datasets, in which the analysis focuses on a sample with no missing (or imputed) outcomes and baseline covariates. In all cases, authors should report their findings and supporting documentation completely and accurately, using information from key *RCT-YES* output tables as shown in this brief.

References and resources

CLEAR—Clearinghouse for Labor Evaluation and Research. U.S. Department of Labor. Accessed at http://clear.dol.gov/about.

"Goal 4: Effectiveness," *Request for Applications, Education Research Grants, CFDA Number: 84.305A* (2016). Institute of Education Sciences, U.S. Department of Education. Accessed at http://ies.ed.gov/funding/pdf/2017_84305A.pdf.

Home Visiting Evidence of Effectiveness. Administration for Children & Families, U.S. Department of Health & Human Services. Accessed at http://homvee.acf.hhs.gov/.

R Project for Statistical Computing. Accessed at https://www.r-project.org/.

RCT-YES download page. Accessed at https://www.rct-yes.com/Home/Download.

RCT-YES home page. Accessed at https://www.rct-yes.com/.

RCT-YES User's Manual. Schochet, P. Z. (2016). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Accessed at https://www.rct-yes.com/Content/PDF/RCT-YES%20Users%20Manual.pdf.

STATA Data Analysis and Statistical Software. Accessed at https://www.rct-yes.com/Home/Download.

Statistical Theory for the RCT-YES Software: Design-based Causal Inference for RCTs, Second Edition (NCEE 2015–4011). Schochet, P. Z. (2016). Washington, DC: Analytic Technical Assistance and Development, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Accessed at https://ies.ed.gov/ncee/pubs/20154011/pdf/20154011.pdf.

Teen Pregnancy Prevention Resource Center: Evidence-Based Programs, Office of Adolescent Health, U.S. Department of Health & Human Services. Accessed at http://www.hhs.gov/ash/oah/oah-initiatives/teen_pregnancy/db/.

What Works Clearinghouse (WWC) home page. Accessed at http://ies.ed.gov/ncee/wwc/.

What Works Clearinghouse™ Reporting Guide for Study Authors, version 1.0 (2012). What Works Clearinghouse, Institute of Education Sciences, U.S. Department of Education. Accessed at http://ies.ed.gov/ncee/wwc/DocumentSum.aspx?sid=235.

WWC Evidence Review Protocol for Dropout Prevention Interventions, version 3.0 (2014). What Works Clearinghouse, Institute of Education Sciences, U.S. Department of Education. Accessed at http://ies.ed.gov/ncee/wwc/documentsum.aspx?sid=25.

WWC Procedures and Standards Handbook, version 3.0 (2014). What Works Clearinghouse, Institute of Education Sciences, U.S. Department of Education. Accessed at http://ies.ed.gov/ncee/wwc/DocumentSum.aspx?sid=19.

WWC Standards Brief – Attrition Standard (2015). What Works Clearinghouse, Institute of Education Sciences, U.S. Department of Education. Accessed at http://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_brief_attrition_080715.pdf.

WWC Standards Brief – Baseline Equivalence (2015). What Works Clearinghouse, Institute of Education Sciences, U.S. Department of Education. Accessed at http://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_brief_baseline_080715.pdf.