Causal Validity Considerations for Including High Quality Non-Experimental Evidence in Systematic Reviews

OPRE REPORT #2018-63
DEKE, MATHEMATICA POLICY RESEARCH
JUNE 2018

OVERVIEW

Federally funded systematic reviews of research evidence play a central role in efforts to base policy decisions on evidence. These evidence reviews seek to assist decision makers by rating the quality of and summarizing the findings from research evidence. Historically, evidence reviews have reserved the highest ratings of quality for studies that employ experimental designs, namely randomized control trials (RCTs). The RCT is considered the "gold standard" of research evidence because randomization ensures that the only thing that could cause a difference in outcomes between the treatment and control groups is the intervention program.[1] However, not all intervention programs can be evaluated using an RCT. To develop an evidence base for those programs, non-experimental study designs may need to be used.

In recent years, standards for some federally funded evidence reviews (i.e., the Home Visiting Evidence of Effectiveness Review [HomVEE] and the What Works Clearinghouse [WWC]) have been expanded to include two non-experimental designs—the regression discontinuity design (RDD) and single case design (SCD). Through the lens of these two reviews, this brief identifies key considerations for systematically and reliably assessing the causal validity of non-experimental studies. Specifically, this brief:

1. Defines causal validity,
2. Provides examples of threats to causal validity and methods that can be used to address those threats,
3. Discusses causal validity ratings in HomVEE and WWC, and
4. Summarizes key considerations for developing standards to assess the causal validity of non-experimental designs.

Dr. John Deke is a senior researcher at Mathematica and co chair of Mathematica's Methods Initiative, with experience in the design and analysis of data from rigorous evaluations based on diverse designs, including random assignment, regression discontinuity, and matched comparison group designs. He currently serves as a principal investigator on several studies and is focused on improving methods for producing and appropriately interpreting research findings.

BACKGROUND ON HomVEE AND WWC EVIDENCE RATINGS

Both HomVEE and the WWC were created to assess research evidence regarding the effectiveness of intervention programs. The Department of Health and Human Services created HomVEE to provide "an assessment of the evidence of effectiveness for home visiting program models that target families with pregnant women and children from birth to kindergarten entry (that is, up through age 5)."[2] Similarly, the Department of Education created the WWC to "provide educators with the information they need to make evidence-based decisions."[3] Both HomVEE and WWC (1) find studies that evaluate the effects of intervention programs in a specified area, (2) rate the quality of each study, and (3) summarize the evidence of effectiveness from each intervention program using only evidence from studies of sufficiently high quality.

One of the core functions of the systematic reviews is to rate the *causal validity* of research findings regarding the effectiveness of evaluated programs. An impact estimate is causally valid if it is expected to reflect a true effect of the program for the individuals included in the study.[4]

[1] The term program is used throughout this brief for simplicity, as programs are often what is evaluated through systematic reviews; however, the methods discussed are relevant to other intervention types (e.g., policy changes, system changes).

[2] See https://homvee.acf.hhs.gov/.

[3] See https://ies.ed.gov/ncee/wwc/.

[4] A closely related concept is "internal validity." The difference between internal validity and causal validity is that causal validity applies to the analysis sample, whereas internal validity applies to the sample that was initially randomly selected or randomly assigned. The impact for the analysis sample could be different from the impact for the initially selected or assigned sample if the characteristics of these two samples differ systematically. However, the impact for the analysis sample could still be causally valid if there are no systematic baseline differences between treatment and comparison group members of the analytic sample.

While causal validity is not the only measure of the usefulness of research evidence, it is the foundation upon which useful evidence is built. The value of research evidence is also affected by other considerations, including external validity, program implementation fidelity, and the validity of outcome measures. Although not the focus of this brief, those other considerations also play an important role in systematic reviews of evidence.

Prior to 2010, only evidence from RCTs had the potential to attain the highest causal validity ratings from the HomVEE and WWC reviews. In an RCT, researchers randomly select which study participants will be offered an opportunity to participate in an intervention program. Randomization ensures that the only *systematic* difference between people offered the intervention program (the treatment group) and the people not offered the intervention program (the control group) is the offer of the program. This means that differences in outcomes between the treatment and control group are due either to random chance or a true effect of the program. The more RCTs that show a positive effect, the more evidence there is that the program is truly effective.

Since 2010, it has also been possible for two non-experimental study designs to be rated as highly as RCTs: Regression Discontinuity Designs (RDDs) and Single Case Designs (SCDs). These designs are highly regarded by methodologists because, as in RCTs, the mechanisms by which individuals are assigned to treatment are observed and well understood by the researcher. Specifically:

- In RDDs, the treatment and comparison groups are determined using a cutoff on a continuous assignment variable. The impact is estimated as the difference in outcomes between individuals just below and just above the cutoff value. With RDDs, the treatment and comparison groups are systematically different only with respect to the assignment variable and that difference can be accounted for through regression adjustment.

- In SCDs, each individual in the study—that is, each *case*—serves as its own control for the purpose of comparison. The researcher controls when each case participates in the program and when they do not. Because the timing of program participation is under the control of the researcher, program participation is not systematically related to other factors that might confound impact estimation.

Developing standards for these two designs has made it possible to expand the base of high quality evidence in contexts where RCTs are often not feasible. For example, SCDs are often used to evaluate interventions in contexts where sample sizes are very small and where it is feasible to switch an intervention on and off relatively frequently to observe changes in outcomes. In contrast, RDDs can be used to study interventions in contexts where the intervention must (for legal or ethical reasons) be allocated according to need, not randomly.

CAUSAL VALIDITY

We define an impact estimate of an intervention program to be “causally valid” if it is *expected* to be truly due to the program, rather than to any other non-random factor. In other words, an impact estimate is causally valid if it is not affected by systematic errors. There can, however, be random errors in an impact estimate. More formally, we follow the potential outcomes framework of Rubin (1974) and define causal validity as equivalence in expectation between a program and comparison group with respect to all factors other than program receipt. This framework was dubbed the Rubin Causal Model (RCM) by Holland (1986). Some of the more common non-experimental study designs that fit within the RCM framework include: matched comparison groups (Rosenbaum and Rubin 1983); difference-in-difference (Imbens and Wooldridge 2009); regression discontinuity (Hahn, Todd, and Van der Klaauw 2001); and instrumental variables (Angrist, Imbens, and Rubin 1996). Of these designs, WWC and HomVEE only have developed standards for RDDs to date.

Developing standards for RDDs and SCDs has made it possible to expand the base of high quality evidence in contexts where RCTs are often not feasible.

Causal validity is necessary for research evidence to productively inform decision making, but it is not sufficient. Any single causally valid impact estimate is not necessarily accurate; the estimate could be affected by random errors. Furthermore, a causally valid impact estimate, even if it is exactly accurate, does not necessarily mean that an intervention program will have similar effects for populations or in contexts that are different from those in the study. Nevertheless, causal validity is the foundation upon which a useful evidence base is built. Without causal validity, other types of validity have little value.

ASSESSING CAUSAL VALIDITY

Assessing the causal validity of evaluation findings involves two related considerations. First, we consider the factors that threaten the causal validity of a finding. Second, we consider the study design and analysis methods that researchers use to address those threats. Findings from studies that use design and analysis methods that can credibly address the most threats are regarded as having the highest causal validity.

Threats to Causal Validity

Causal validity is threatened by systematic errors in impact estimates. These systematic errors are often referred to simply as "bias." While we can at least partially protect ourselves from *random* errors using interpretive tools such as confidence intervals, the false discovery rate, or Bayesian posterior probabilities (Genovese and Wasserman 2002; Storey 2003), there is much less that we can do to protect ourselves from systematic errors in impact estimates.

There are two broad types of systematic errors: (1) those that diminish as sample size increases and (2) those that do not diminish as sample size increases. Both types of errors can be problematic, but errors that do not diminish as sample size increases pose the greatest threat to causal validity.

Examples of systematic errors that diminish as sample size increases (sometimes called "finite sample bias") include the following:

1. **Errors that arise when estimating the impact of participating in a program when the *offer* of a program is very weakly correlated with participation in the program.** Estimating the impact of participating in a program is sometimes called estimating the impact of treatment on the treated (Bloom 1984) or estimating the complier average causal effect (CACE) (Angrist, Imbens, and Rubin 1996). This type of estimate can be biased if the treatment variable is weakly correlated with program participation. In cases of extremely low correlations between treatment status and program participation, the estimate of the CACE reduces to the Ordinary Least Squares (OLS) estimate—that is, what we would get if we regressed the outcome on program participation without accounting for the random assignment treatment variable at all (Stock and Yogo 2005). Since the OLS estimate is typically biased, this means that the CACE estimate is also biased. The WWC's approach to assessing this issue when reviewing studies is described in WWC (2017).

2. **Errors due to functional form misspecification.** When the effect of an intervention is estimated using a regression analysis, the impact estimate can be systematically incorrect if the following conditions are met: (1) the treatment variable (T) is highly correlated with another covariate (X) in the regression, (2) X is highly correlated with the outcome (Y), and (3) the relationship between X and Y is not correctly specified in the regression equation (for example, the relationship is specified to be linear when in truth it is cubic).[5] This type of error occurs in studies using matched comparison group designs or regression discontinuity designs.

Examples of systematic errors that do not diminish as sample size increases include the following:

1. **Errors due to self-selection into program participation.** In a study that calculates the program impact as the difference in outcomes between individuals who *choose* to participate in a program and individuals who *choose not* to participate in the program, the estimated impact could be due to the preexisting difference in individuals that led them to make different program participation choices. For example, individuals who choose to participate in a program might be more motivated, or better able, to make a change in their lives than people who choose not to participate. That higher motivation—which existed before program participation—might lead to better outcomes than less motivated individuals in the comparison group.

2. **Errors due to non-random missing outcome data.** Systematic errors in impact estimates could arise when outcome data is missing in a way that is related both to treatment status and outcomes. For example, in a study of financial incentives for teachers whose students show the highest performance gains, teachers in the treatment group might have an incentive to discourage low-ability students from taking the test used to measure the teacher's performance. Impacts calculated using the available outcome data would overestimate the effect of the financial incentives on student test scores.

[5] It is still better to include X than to exclude it—errors associated from ignoring an important covariate altogether are much worse than functional form misspecification errors.

3. **Errors due to confounding factors.** Systematic errors in impact estimates could arise when everyone exposed to the treatment is also exposed to some other factor that also affects outcomes. For example, if everyone in the treatment group is in one state and everyone in the comparison group is in another state, then the effects of the treatment are confounded with the effects of living in one state versus the other.

In addition to the examples listed here, there can also be threats to validity that are unique to specific designs. The design-specific threats may be related to the general categories listed above, but they may need special consideration in the context of specific designs. For example, in RDDs a threat that requires special attention is manipulation of the assignment variable, which is essentially a special case of self-selection bias in which assignment variable values are "faked" in order to change the assignment status of some individuals.

Methods to Address Threats to Causal Validity

A variety of methods exist to address threats to causal validity. Some methods work by adjusting an impact estimate to correct for observed threats, while other methods attempt to address threats that cannot be directly observed (either by avoiding the threats or bounding them).

The following are examples of methods used to address threats to causal validity that HomVEE and WWC standards take into account:

1. **Preventing threats by design.** The most compelling method to address threats to causal validity is to design a study so that a threat is either impossible or highly unlikely (so long as the study is conducted in accordance with the design). For example, studies based on a randomized design, regression discontinuity design, or single case design all prevent self-selection bias by design.

2. **Regression adjustment for observed differences.** Systematic differences in observed characteristics between a program and comparison group, possibly due to self-selection or missing outcome data, can be mitigated using regression adjustment. However, even if these methods are successful, unobserved differences could remain (leading to a biased treatment effect).

3. **Matching or weighting methods to establish baseline equivalence.** This approach can be used to construct a program and comparison group that are equivalent with respect to observed characteristics at baseline. This approach can be used as an alternative to regression adjustment in order to adjust for observed differences between program participants and non-participants. It can also be used in conjunction with regression adjustment to reduce bias due to functional form misspecification.

4. **Model-based bounding.** Some forms of bias cannot be directly observed, but the magnitude of the bias can potentially be bounded. The bounding approach used by the WWC and HomVEE is to construct bounds that work under scenarios that are deemed realistic based on a model of bias in which the selected parameter values of the model are informed by data and theoretical considerations. This model-based approach is the basis of the attrition standard used by the WWC and HomVEE (Deke and Chiang 2016).[6]

Examples of approaches to address threats to causal validity

Approach	Threats addressed
Prevent threats by design	Potentially all threats can be reduced or possibly even eliminated at the design stage of a study.
Regression adjustment for observed differences	Systematic differences between the treatment and comparison group with respect to observable characteristics of study participants. Such differences could arise from threats like self-selection into treatment or missing outcome data.
Matching or weighting methods	Misspecification of the functional form of a regression; also the same threats addressed by regression adjustment.
Bounding	Systematic differences between the treatment and comparison group with respect to *unobserved* characteristics of study participants due to, for example, missing outcome data (attrition).

[6] Another approach is to construct bounds that are guaranteed to include the true effect under worst case scenarios (Manski 1990; Horowitz and Manski 1995; Lee 2009). While those approaches yield highly credible bounds because they require minimal assumptions, the bounds are often so wide that they provide little useful information.

In addition to the examples listed here, there can also be approaches that are unique to specific designs. For example, in RDDs, bias due to functional form misspecification can be addressed using the approach of estimating the impact regression within a bandwidth around the cutoff on the assignment variable. In a sense, this is a special case of weighting or matching to achieve balance, but the method is sufficiently customized for the RDD context that it is best considered as a distinct method.

RATING CAUSAL VALIDITY

To rate the causal validity of evaluation findings, evidence reviews need to be able to systematically and reliably assess the likely magnitude of the threats described above and the likely effectiveness of the methods used to address those threats. A systematic and reliable approach to that assessment is needed to support an impartial and transparent review process. In other words, reviewing evidence in this context needs to be more science than art. By contrast, a review process that depends on subjective assessments that are inconsistent across reviewers (for example, the review process used to referee submissions to academic journals) would not be sufficient for systematic reviews.

To facilitate a systematic and reliable assessment, evidence reviews use standards-based procedures that can be consistently implemented by trained reviewers. Examples of evidence standards used by the WWC and HomVEE include the attrition standard, the baseline equivalence standard, regression discontinuity design standards, and single case design standards (WWC 2017). These standards specify objective criteria that can be implemented consistently by trained reviewers to assess the causal validity of study findings.

HomVEE and the WWC have three ratings for the causal validity of evidence: *high*, *moderate*, and *low.*[7]

HomVEE. The current ratings categories and criteria for the HomVEE review are summarized in Table 2.[8] Studies receive these ratings as the result of standards-based reviews. Prior to the development of standards for RDDs and SCDs, only RCTs could receive a *high* rating, and only if they had low attrition. Studies with a *moderate* rating were either RCTs with high attrition or studies based on a matched comparison group design. In both cases, the study had to demonstrate the equivalence of the treatment and comparison groups with respect to key covariates. Studies with a *low* rating were studies that failed to demonstrate equivalence with respect to key covariates.

WWC. In 2010 the WWC released pilot standards for SCDs and RDDs—two non-experimental designs that, like RCTs, can receive a high rating. The decision to add these two designs to the same rating category as RCTs was influenced by several factors.

First, methodologists demonstrated that when well executed, studies using these designs have the potential to achieve a level of causal validity that is much closer to a randomized experiment than to other non-experimental designs (Shadish, Cook, and Campbell 2002; Cook and Steiner 2009). Second, these designs can sometimes be used in contexts where randomized experiments are not feasible. By recognizing the high-quality evidence that these designs can produce, it may be possible to produce high quality findings in a broader range of contexts, leading to more high quality information available to decision makers. Third, it was possible to develop standards with objective criteria that enabled systematic, reliable reviews of studies that use these designs.

CONCLUSION

Including studies using non-experimental designs that are not currently eligible to receive high ratings requires the development of new standards that can be used by trained reviewers to assess the causal validity of study findings. Using the standards, reviewers would need to systematically and reliably:

1. Assess all relevant threats to the causal validity of the study's findings. The threats to causal validity described in this brief are examples of the types of threats that reviewers would need to assess. However, in the process of developing new standards additional threats that are unique to a given design may need to be considered. For example, when the WWC developed the RDD standards it was determined that manipulation of the RDD assignment variable is a threat to causal validity that is unique to RDD.

2. Assess the likely efficacy of methods used to address those threats. The methods used to address

[7] The WWC uses the terms meets standards, meets standards with reservations, and does not meet standards.

[8] This table is based on http://homvee.acf.hhs.gov/Review-Process/4/Producing-Study-Ratings/19/5.

Table 2. Summary of Study Rating Criteria for the HomVEE Review

	HomVEE Study Rating	
	High	Moderate
Randomized Controlled Trials	Random assignment Meets WWC standards for acceptable rates of overall and differential attrition No reassignment; analysis must be based on original assignment to study arms No confounding factors; must have at least two participants in each study arm and no systematic differences in data collection methods Baseline equivalence established on tested outcomes and demographic characteristics OR controls for these measures	Reassignment OR unacceptable rates of overall or differential attrition Baseline equivalence established on tested outcomes and demographic characteristics **AND** controls for baseline measures of outcomes, if applicable No confounding factors; must have at least two participants in each study arm and no systematic differences in data collection methods
Regression Discontinuity	Integrity of assignment variable is maintained Meets WWC standards for low overall and differential attrition The relationship between the outcome and the assignment variable is continuous Meets WWC standards for functional from and bandwidth	Integrity of assignment variable is maintained Meets WWC standards for low attrition Meets WWC standards for functional from and bandwidth
Single Case Design	Timing of intervention is systematically manipulated Outcomes meet WWC standards for interassessor agreement At least three attempts to demonstrate an effect At least five data points in relevant phases	Timing of intervention is systematically manipulated Outcomes meet WWC standards for interassessor agreement At least three attempts to demonstrate an effect At least three data points in relevant phases
Matched Comparison Group	Not applicable	Baseline equivalence established on tested outcomes and demographic characteristics **AND** controls for baseline measures of outcomes, if applicable No confounding factors; must have at least two participants in each study arm and no systematic differences in data collection methods
Studies receive a low rating if they do not meet the requirements for a high or moderate rating.		

threats to causal validity described in this brief are examples, but new designs are likely to include new methods that would require new standards to assess their efficacy. For example, causal validity in RDD studies is especially dependent on correctly specifying the functional form of the regression used to estimate impacts. A unique standard for RDD studies had to be developed to address this issue.

3. **Use those assessments to rate the causal validity of the study's findings.** Standards for reviewers need to specify how to classify a study. A critical consideration when developing new standards is to ensure that all findings, regardless of study design, that receive the same rating have approximately the same level of causal validity. For example, studies based on RDD or SCD must satisfy many more criteria than RCTs because those designs face more threats to validity and require the use of more methods to address those threats than an RCT. But if a study based on an RDD or SCD does attain the highest quality rating, it is correct to conclude that its causal validity is closer to that of an RCT than it is to studies that receive a moderate rating.

REFERENCES

Angrist, J.D., Imbens, G.W., and Rubin, D.B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*, 444–472.

Bloom, H. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review*, *8*(2), 225–246.

Cook, T.D., and P.M. Steiner. (2009). Some empirically viable alternatives to the randomized experiment. *Journal of Policy Analysis and Management, 28*(1), 165–166.

Deke, J., and Chiang, H. (2016). The WWC Attrition Standard: Sensitivity to Assumptions and Opportunities for Refining and Adapting to New Contexts. *Evaluation Review, 41*(2), 130-154.

Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. Roy. Statist. Soc. Ser. B* 64, 499–517.

Hahn, J. Todd, P. and Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, *69*(1), 201–209.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*, 945–970.

Horowitz, J. L. and Manski, C. F. (1995). Identification and robustness with contaminated and corrupted data. Econometrica, 63, 281–302.

Imbens, G. and Wooldridge, J. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*.

Lee, D. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies, 76*(3), 1071–1102.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *American Economic Review, Papers and Proceedings*, *80*, 319–323.

Rosenbaum, Paul R., and Donald B. Rubin. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66,* 688–701.

Shadish, W.R., T.D. Cook, and D.T. Campbell. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

St. Clair, T., Hallberg, K., and Cook, T.D. (2016). The validity and precision of the comparative interrupted time-series design: Three within-study comparisons. *Journal of Educational and Behavioral Statistics, 41*(3), 269–299.

Stock, J. H., and M. Yogo. (2005). “Testing for Weak Instruments in Linear IV Regression.” In D. W. K. Andrews and J. H. Stock (eds.), *Identification and Inference for Econometric Models: A Festschrift in Honor of Thomas J. Rothenberg.* Cambridge, UK: Cambridge University Press.

Storey, John D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *The Annals of Statistics*, *31*.6, 2013–2035.

What Works Clearinghouse. (2017). *WWC standards handbook (version 4.0).* Washington, DC: U.S. Department of Education, Institute of Education Sciences. Retrieved February 26, 2018, from https://ies.ed.gov/ncee/wwc/handbooks.

This brief was prepared for the Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services. It was developed under Contract Number HHSP23320095651WC. RTI International assisted with preparing and formatting the brief. The ACF project officers are Anna Solmeyer and Laura Nerenberg. The RTI project director is Anupa Bir.

Suggested citation: Deke, John (2018). *Causal validity considerations for including high quality non experimental evidence in systematic reviews* (OPRE Report 2018 63), Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

This brief and other reports sponsored by the Office of Planning, Research, and Evaluation are available at www.acf.hhs.gov/opre.

Disclaimer: The views expressed in this publication do not necessarily reflect the views or policies of the Office of Planning, Research, and Evaluation, the Administration for Children and Families, or the U.S. Department of Health and Human Services.