CLEAR CAUSAL EVIDENCE GUIDELINES, VERSION 2.1

The Clearinghouse for Labor Evaluation and Research (CLEAR) identifies and summarizes many types of research, including descriptive, implementation, and impact studies. For causal research—defined as research intended to assess the effectiveness or impact of a program, policy, or activity (hereafter referred to as an "intervention")—CLEAR provides an objective assessment and rating of the degree to which the research establishes the causal impact of the intervention on the outcomes of interest. This document describes CLEAR's evidence guidelines for rating the quality of this kind of research. The guidelines are sorted by research design, with regression designs (including instrumental variables) in Section A, randomized controlled trials (RCTs) in Section B, and interrupted time series (ITS) studies in Section C. The final two sections of the document describe some considerations for all research designs, as well as how the guidelines evolved and how CLEAR handles design concerns that are not specified in the current version of the guidelines. A companion addendum provides additional examples of how to apply these guidelines.[1]

CLEAR has three ratings to describe the strength of the causal evidence in a study: high, moderate, and low.

- A **high rating** means we are confident that the estimated effects are solely attributable to the intervention examined. Two types of studies can receive a high rating: (1) well-conducted RCTs that have low attrition and no other threats to study validity and (2) ITS designs with sufficient replication wherein the intervention condition is intentionally manipulated by the researcher (Table 1).[2] RCTs and ITS designs that do not qualify for a high rating can be evaluated against CLEAR's evidence guidelines for regression analyses.
- A **moderate rating** means we are somewhat confident that the estimated effects are attributable to the intervention studied, but there might be other contributing factors that were not included in the analysis. Research that meets the CLEAR guidelines for regression designs receives a moderate rating; this includes RCTs and ITS designs that do not receive a high rating.
- Research that does not meet the criteria for a high or moderate rating receives a **low rating**, which indicates that we cannot be confident that the estimated effects are attributable to the intervention studied. Other factors likely contributed to the estimated effects.

A high rating does not necessarily mean the study showed positive impacts, only that the analysis meets high methodological standards, and the causal impacts estimated (in any direction) are credible. Similarly, a low rating does not mean the study's results are not useful for some purposes, but they should be interpreted with caution. Ratings of causal evidence only reflect the extent to which a given

[1] These guidelines are for comprehensive profile reviews, as described in the *CLEAR Policies and Procedures*. These are more detailed than the brief highlights reviews; highlights reviews do not assess the quality of the causal evidence that a piece of research presents.

[2] Research shows that ITS designs can provide strong causal evidence (see Shadish, W., T. Cook, and D. Campbell. (2002). "Quasi-Experiments: Interrupted Time-Series Designs." In *Experimental and Quasi-Experimental Designs for General Causal Inference*. Boston: Houghton Mifflin Company, 171–206). ITS designs can also be seen as a hybrid of single-case and regression discontinuity designs, both of which have been judged by experts to provide strong causal evidence when well-executed (see the *WWC Procedures and Standards Handbook,* version 3.0, available at http://ies.ed.gov/ncee/wwc/documentsum.aspx?sid=19). However, CLEAR leadership anticipates that ITS designs in topic areas of interest to CLEAR will rarely be strong enough to receive a high causal evidence rating.

study shows a causal effect (internal validity), not the extent to which that causal effect would be expected in different contexts (external validity).

Table 1. CLEAR evidence guidelines and highest possible ratings, by study design

Study design	Applicable CLEAR guidelines	Highest possible causal evidence rating
RCT	RCT	High
ITS	ITS	High
Matched comparison group	Regression	Moderate
Difference-in-differences	Regression	Moderate
Fixed effects (group or individual)	Regression, with special criteria	Moderate
Instrumental variables (including two-stage least squares, the Heckman two-step correction, and limited information maximum likelihood)	Regression, with special criteria	Moderate
Other regression methods (including ordinary least squares, hazard, logit, probit, and tobit)	Regression	Moderate
Correlational or descriptive studies that make causal claims	Regression	Low
Pre/post (for example: ANOVA, t-test)	ITS	Low
Noncausal (implementation studies, correlational or descriptive studies that do not make causal claims)	Implementation or descriptive	Not applicable

A. Regression analyses

Regression analysis is a statistical method that can be used to calculate the effect of an intervention on an outcome, isolating this effect from any other factors that also could affect the outcome (such as educational attainment and work history). This kind of analysis can involve several techniques, including ordinary least squares, logit (also called logistic regression), probit, tobit, matching methods, and hazard models. CLEAR's guidelines provide criteria for rating studies that use these methods as well as additional criteria for special applications of regression techniques, including fixed effects, random effects, difference-in-differences, and instrumental variables models.

Fixed effects are components of statistical analysis models that account for unobserved, time-invariant characteristics of sample members that might affect (1) whether they received the intervention and (2) the outcomes of interest. For example, at the individual level, a sample member may choose to participate in a job training program because of some unobserved personal characteristic—such as motivation and cognitive ability—which could in turn affect that person's earnings-related outcomes. If the study authors fail to take statistically adjust for these factors, the results of their analysis would be biased. In this example, the statistical model could include *individual fixed effects* to account for all the participants' time-invariant characteristics (such as motivation).

Other types of fixed-effects models focus on group-level interventions—meaning that people could have been affected by the intervention because they were in a group (such as a state or a firm) that was subject to it, without having opted to participate. In these models, *group fixed effects* account for unobserved, time-invariant characteristics of the groups that might affect both the receipt of the intervention and the outcomes of interest. For instance, consider an analysis of the effect of minimum wage laws on the earnings of workers in a state. If the state chooses to adopt the law, the workers in

that state are subject to it. An analysis of individual workers' earnings might include a state-level fixed effect that would hold constant other factors in the state, such as a strong union presence, that might affect both the decision to adopt the minimum wage law and workers' earnings.

Another regression technique is the use of *random-effects* models. These models are similar to fixed-effects models in that they models an individual-specific effect. However, random-effects models rely on the assumption that time-invariant, unobserved characteristics are not correlated with other explanatory variables in the model.

A difference-in-differences model is a special kind of fixed-effects model. Difference-in-differences models show the changes in outcomes over time for the group receiving the intervention versus the changes over the same period for a comparison group that did not receive the intervention (or received a different one).[3] These models are often used to assess an intervention adopted at a group level, such as a policy at the state level, and the analysis also takes place at the group level. For example, suppose states adopted a policy to increase the minimum wage over a 10-year period, with some adopting it in each year and some never adopting it, and a researcher analyzed the effect of the law on the states' unemployment rates. The researcher would use a difference-in-differences model to compare the changes in unemployment rates in states that adopted the minimum wage policy versus states that did not adopt the policy. Using this approach, the researcher can account for changes in the outcome variable that would have occurred over time for reasons that are not related to the policy, as well as for "fixed" differences between the intervention and comparison groups that are unrelated to receiving the intervention.[4]

1. Criteria for all regression models

CLEAR uses the following criteria to evaluate the causal validity of all studies that involve regression models. All such studies must meet the Regression.1 through Regression.3 criteria to receive a moderate causal evidence rating; failure to meet one or more of these criteria results in a low causal evidence rating. Furthermore, studies in which group-level effects are estimated must meet Regression.4, and studies that involve use of random effects must meet Regression.5 to receive a moderate rating.

Criterion Regression.1: Were the intervention and comparison groups similar before the intervention? The intervention and comparison groups being analyzed must be similar before the intervention begins. This ensures that the two groups are comparable and that the experiences of the comparison group present a valid picture of what would have happened to the intervention group if it had not been exposed to the intervention. Two types of comparability are relevant for determining causal validity: comparability on observed and unobserved characteristics.

Observed characteristics. Comparability on observed characteristics means that the two groups being analyzed are similar on key baseline characteristics, or the study authors have adjusted for differences between the groups by including appropriate controls in the regression. In a cross-sectional regression,

[3] Difference-in-differences models can also be described as a "short ITS" design with a comparison group, or as a comparison group design with pre- and post-intervention data. CLEAR definesdifference-in-differences analyses as those that compare changes over time in intervention group outcomes with those of an explicit comparison group consisting of units other than those treated; thus, they are subject to the Regression guidelines. In contrast, ITS analyses are used to compare outcomes over time within a series of observations on the same units *without* an external comparison group consisting of different units; these analyses are subject to the ITS guidelines. Reviewers carefully determine which standards to use after considering the design and analysis approach, consulting CLEAR leadership as necessary.

[4] Designs in which people receive an intervention, then stop receiving it, and then receive it again present a separate set of methodological issues that are outside the current scope of these guidelines.

to establish comparability between the groups being analyzed on observed characteristics, the authors could compare characteristics measured before the intervention for the two groups and show that the differences between them are not statistically significant (that is, $p >= 0.05$ in a two-tailed test). If the authors do not display the results of this comparison, or if they show that groups do have statistically significant differences, then the authors must also control for these characteristics in the analysis. Typically, basic demographic information alone cannot establish comparability of the groups or serve as a sufficient control in a cross-sectional regression; pre-intervention or lagged values of the key outcome measure will usually be needed. Use of a gain score as the dependent variable does not satisfy this criterion. In some cases, reviewers may be concerned that the differences between the groups, although not statistically significant, are too large to effectively control for; the principal investigator (PI) will seek guidance from the CLEAR leadership team in such cases.

CLEAR reviewers examine the control variables and lags in the pre-intervention outcome included in the analysis. The requirements for the number of lags and the period they cover, as well as for types of control variables, in order for a cross-sectional regression to meet this criterion vary by the topic area and outcome being examined, and are specified in the review protocol for each topic area. For example, the specified pre-intervention characteristics for research that analyzes the employment outcomes of youth for the Opportunities for Youth topic area include pre-intervention measures of employment (lagged employment variables), age, gender, race/ethnicity, and geographic location. A study may present comparisons for other characteristics at baseline that are not required by the relevant topic-area protocol to establish comparability. But if a study shows sizeable or statistically significant differences on these variables and they are not included as controls, the study may not be able to receive a moderate causal evidence. Reviewers indicate their concerns about these types of the differences to the PI, who then seeks guidance from the CLEAR leadership team.

For studies using panel data, *including difference-in-differences models and models with individual fixed effects*, authors must demonstrate equivalent *trends* between the intervention and comparison groups being analyzed to satisfy this criterion (the required variables to examine vary by topic area and outcome and are specified in the review protocol). That is, if an author identifies the effects of an intervention by pointing to changes in an outcome over time, then the changes in that outcome before the intervention should be the same across the intervention and comparison groups (but the levels of pre-intervention outcomes need not be the same for the two groups). Authors can demonstrate equivalent trends by inspection or, in the case of only one pre-intervention period, the use of placebo tests (see the addendum for examples). If the authors do not attempt to show equivalence by one of these methods, or if the trends do appear to differ, then they must adequately control for time-varying characteristics that might affect the outcomes. In some cases, the differences in trends might be too large to effectively control for. Reviewers indicate their concerns about these types of the differences to the PI, who then seeks guidance from the CLEAR leadership team.

Unobserved characteristics. For a study to meet the Regression.1 criterion, the intervention and comparison groups also need to be comparable on unobserved characteristics. This guards against situations in which the intervention and comparison groups appear to be similar on observed characteristics, but there is an obvious selection mechanism, documented in the study, whereby people enter the intervention group based on an unobserved characteristic, and it affects both the decision to participate and the outcome of interest. For example, suppose that community college students who register for classes on time—that is, before the registration deadline—are eligible to participate in an intervention, but the comparison group consists of students who registered late. Those who registered on time likely have some innate characteristic, such as motivation or ambition, that those who registered late do not have, and this trait would affect both participation in the program and the post-intervention outcomes. Therefore, using the late registrants as a comparison group does not

provide a valid picture of how the program participants would have fared in the absence of the intervention.

Typically, any intervention triggered by changes in the outcome variable will likely have issues related to noncomparability of unobservable characteristics. More generally, if reviewers identify a plausible and documented selection mechanism that is not controlled for in the analysis, they indicate their concerns to the PI, who then seeks guidance from the CLEAR leadership team.

Criterion Regression.2: Were there no confounding factors? Except for the intervention, the conditions for the comparison group should be the same as those experienced by the intervention group. One somewhat common counfound is known as the N=1 confound, when all cases from one study arm were in one state, county, local workforce investment area, school, or cohort. In such cases the effect of the intervention cannot be disentangled from the effect of the single unit in which the intervention was delievered. For instance, suppose that an alternative sentencing program for low-level drug offenders were implemented in one county, and the outcomes of people sentenced through the alternative sentencing program were compared with those of people with similar offenses but served through traditional sentencing in nearby counties. Because only one county implemented the alternative sentencing, we cannot disentangle the effects of the alternative sentencing program from the effect of the county. As one example, there could be other criminal justice interventions implemented simultaneously in the county, and they—not the alternative sentencing—could be responsible for the observed effects. Confounding factors can also include time-varying factors that differentially affect the intervention group. For instance, in the case of state policy variation, other state policies occurring over the same time period could also affect the outcome of interest.

If reviewers identify a potential N=1 problem, they determine whether there was variation on another dimension that can be used to assess whether the study meets criteria for a differences in differences analysis. For example, if the treatment was implemented in only one county, but the analysis included data for multiple time points, the analysis could be reviewed using the Regression.1 criteria for panel data.

Criterion Regression.3: Was sample members' anticipation of the intervention either unlikely or appropriately controlled for? Intervention group members sometimes adjust their behavior in anticipation of participating in a new program or becoming subject to a new policy. For example, suppose a new state safety standard was announced and would go into effect in six months. During the six months between this announcement and enforcement of the new standard, businesses might begin increasing their safety precautions in anticipation of the new law, potentially reducing their rate of workplace injuries. Therefore, an analysis of the effect of the law on workplace injuries could not include the six-month lagged injury rates as controls in the regression model, as including these rates would not accurately reflect the law's true impact. In most studies, anticipating the intervention will not be possible because of the nature of the intervention and participation in it. However, when it's possible for treated subjects to anticipate the intervention, the study authors must address it effectively to meet this criterion (in the example above, this would mean disregarding the six months of data between the announcement and enforcement).

2. Special criterion for estimates of group-level effects

Some research designs include group-level (rather than individual-level) fixed effects or group-level control variables to account for pre-intervention lags in the outcome variable (for instance, state-level average earnings from the Current Population Survey before changes are made to the minimum wage law). These designs must satisfy an additional criterion.

Criterion Regression.4: Were there no changes in group composition? The composition of the intervention and comparison groups should not change in ways related to the outcome of interest. For example, changes in minimum wage laws could induce some workers to leave certain states and enter neighboring ones. If this selective migration were substantial enough, the direct effects of the minimum wage law could become conflated with the effect of changes in the composition of the labor force. In this case, the study must provide evidence that there was not substantial selective migration into or out of the states affected by the policy change.

CLEAR uses conservative and liberal standards for acceptable levels of migration (defining these levels the same way as the attrition boundaries established by the Institute of Education Sciences' What Works Clearinghouse [WWC] and described under Criterion RCT.3). In each topic area that might include reviews of group-level effects, the review protocol specifies the migration standard to use for that topic area. Reviewers apply the attrition thresholds shown in Table 2 (see Section C) when assessing overall and differential migration, substituting the differential migration rate for the differential attrition rate and the overall migration rate for the overall attrition rate.

CLEAR reviewers use the conservative standards when there is reason to believe that relatively more of the migration might be caused by the intervention examined. For example, suppose a study estimated the impact of a generous tuition assistance program on employee retention, using firm-level data. Because people might be induced to apply to or leave a firm based on the benefits they receive, the conservative migration standard would be used to assess this study. In contrast, reviewers use the liberal standards when the migration seems to be less likely caused by the intervention examined. For example, suppose a study estimated the impact of auto-enrollment policies for 401(k) plans on employee savings rates, using firm-level data. Because such policies can have large impacts on savings without affecting people's employment decisions, the liberal migration standard would be appropriate.

When migration is shown to be within the migration cutoffs, the study author does not have to make additional adjustments. However, if migration exceeds the cutoffs, the author must use data on individual or group characteristics to account for measurable changes in composition.

The topic-area protocols list the types of group-level analyses that do not need to meet this criterion. For example, in an industry-level analysis of the impact of a safety policy on injury rates, the least safe companies might go out of business after the policy is implemented, implying changes in industry composition. In this example, the compositional change could be considered part of the impact of the policy that might lead to changes in injury rates. Thus, the reviewer would waive Criterion Regression.4 for industry-level analyses of this type.

3. Special criterion for random effects

Besides Criterion Regression.1 through Criterion Regression.3 (and Criterion Regression.4 if the analysis includes groups), random-effects (RE) models must satisfy an additional criterion:

Criterion RE.1: Use of random effects over fixed effects. When the random-effects model is valid, the fixed-effects estimator will still produce unbiased estimates of the relationships of interest, but they will be estimated less efficiently than the random-effects estimates. However, if unobserved characteristics are correlated with explanatory variables in the model, the random-effects estimates will be biased. Therefore, in general, CLEAR prefers the fixed-effects model unless there is compelling evidence that the author has included all time-invariant factors that could be correlated with unobserved characteristics affecting the outcome. If the review team identifies a factor that could be correlated with other explanatory variables but is not included in the random-effects model, the study

does not meet this criterion. In addition, studies using random effects must report a specification test justifying the use of random effects over fixed effects.

4. Special criteria for instrumental variables models

Instrumental variables techniques are often used when a sample member's chance of receiving an intervention is determined by a combination of endogenous and exogenous factors. Endogenous factors are factors related to both receipt of the intervention *and* the outcomes of interest. For example, highly motivated people might be more likely to participate in a job training program but also tend to have better outcomes even without the intervention, so individual motivation is an endogenous factor. Exogenous factors are factors related to receiving the intervention but *not* related to outcomes (after controlling for factors that affect the outcome irrespective of the intervention, such as age or gender). For instance, an exogenous factor could be a lottery to select which program applicants will be admitted to an oversubscribed program.

Receipt of an intervention can be related to unmeasured endogenous factors. If this is the case, a simple regression of outcomes on an indicator of intervention receipt will lead to biased estimates of the causal relationship. For example, people can self-select for a training program, or firms can be selected for monitoring by an enforcement agency via a nonrandom process. A simple analysis that does not account for this selection will inappropriately attribute all observed effects to the program when some effects could instead be due to individual motivation or to the nonrandom selection. One approach to dealing with this problem is to estimate impacts using only exogenous factors that affect whether a person received the intervention, while filtering out the influence of the endogenous factors. These exogenous factors are sometimes called instrumental variables. There are many ways to estimate impacts using instrumental variables, including two-stage least squares, the Heckman two-step correction, and limited information maximum likelihood.

Research using an instrumental variables approach must satisfy Criterion IV.1 and Criterion IV.2 to receive a moderate causal evidence rating (certain instrumental variables designs must also satisfy Criterion IV.3); otherwise, the research receives a low rating. These three criteria subsume the first three regression criteria, but if the study uses group-level estimates, it must also satisfy Criterion Regression.4.

Criterion IV.1: Does the instrument have sufficient strength? The instrument must be strong enough to predict whether a sample member received the intervention; with weaker instruments, impact estimates might be biased. Therefore, to meet this criterion, studies must report the results of a test of an instrument's strength. In one common test, authors use a first-stage equation that models intervention as a function of the instrument and all explanatory variables. The test is based on the first-stage F-statistic for the null hypothesis that the instrument has no effect on intervention. If the F-statistic exceeds 10, the instrument is considered to be sufficiently strong. See the addendum for other ways to test an instrument's strength.

Criterion IV.2: Does the instrument satisfy the exclusion restriction (or instrument exogeneity)? The only way in which the instrument can affect the outcome must be through receipt of the intervention. The study author must make a clear and convincing case that the exclusion restriction is satisfied; otherwise, the reviewer assumes it's not.

In addition, research designs with multiple endogenous variables and instruments must satisfy a third criterion.

Criterion IV.3: Is the rank condition satisfied? There must be at least as many instruments as there are endogenous explanatory variables, and those instruments must not be highly collinear with one another. In technical terms, this is often described as *satisfying the rank condition*. See the addendum for examples of how to test whether this condition is satisfied.

5. Note on matching designs

Matching and weighting designs, including propensity-score matching designs, are evaluated according to the guidelines for regression analyses. A matching analysis must match on all the control variables specified in the topic-area protocol; those that do so need not include these variables in the regression. If the matching analysis does not attempt to match on all key control variables, or if the matching process was unsuccessful on one or more of these variables, the regression analysis must include them. A matching process is determined to be successful if, after matching, the differences between the characteristics of the intervention and comparison groups are not statistically significant (that is, $p >= 0.05$ in a two-tailed test). In addition, matching methods must meet Criterion Regression.1 through Criterion Regression.4, as applicable.

In some instances, authors might choose to apply weights derived from a matching process (for example, propensity-score weights) to study data for analysis. In these cases, the comparability of the groups on observed variables must be demonstrated on the weighted data, in addition to meeting the other applicable regression criteria.

6. Note on imputation

Regression analyses that involve imputing baseline or outcome variables for a portion of the analysis sample can receive a moderate causal evidence rating as long as all other applicable regression criteria have been met. However, the profiles of these studies will include information on the potential interpretation issues that arise when baseline or outcome variables have been imputed.

B. Randomized controlled trials

RCTs must satisfy three criteria to receive a high causal evidence rating. RCTs that fail to meet one or more RCT criteria will be evaluated using the regression guidelines.

Criterion RCT.1: Were there no confounding factors? If random assignment is properly implemented, the only thing that differs between the intervention and comparison groups is the intervention itself. However, RCTs can have confounding factors that make it impossible to separate the effect of the intervention from the effects of other factors. For example, if only one school implemented a schoolwide program for youth, it would be impossible to separate the effect of the program from the effect of the staff and the environment at that school.

Criterion RCT.2: Was sample attrition low? Sample attrition is a key factor in determining the strength of evidence for RCTs. CLEAR considers both the overall sample attrition rate and the difference in sample attrition rates between the intervention and comparison groups because both contribute to the potential bias of the estimated effect of an intervention.

There are conservative and liberal standards for acceptable levels of attrition. The review protocols specify the attrition standard used for each topic area. When the attrition seems to be more endogenous (rather than exogenous) to the intervention—for example, disadvantaged youth choosing whether to participate in a residential career training program—the conservative standards are applied.

When the attrition seems more exogenous to the intervention—for example, employers cutting back the number of slots in a training program because of reduced funding—the liberal standards are applied.

Attrition rates are based on the number of cases in the analysis sample with measured (as opposed to imputed) values of the outcome measures. For a given level of overall attrition, Table 2 presents the maximum differential attrition rate between the intervention and comparison groups that is acceptable. The higher the rate of overall attrition, the lower the rate of differential attrition must be to be considered acceptable.

Table 2. Thresholds of acceptable combinations of overall and differential attrition (percentages)

	Differential attrition			Differential attrition	
Overall attrition	Conservative boundary	Liberal boundary	Overall attrition	Conservative boundary	Liberal boundary
0	5.7	10.0	34	3.5	7.4
1	5.8	10.1	35	3.3	7.2
2	5.9	10.2	36	3.2	7.0
3	5.9	10.3	37	3.1	6.7
4	6.0	10.4	38	2.9	6.5
5	6.1	10.5	39	2.8	6.3
6	6.2	10.7	40	2.6	6.0
7	6.3	10.8	41	2.5	5.8
8	6.3	10.9	42	2.3	5.6
9	6.3	10.9	43	2.1	5.3
10	6.3	10.9	44	2.0	5.1
11	6.2	10.9	45	1.8	4.9
12	6.2	10.9	46	1.6	4.6
13	6.1	10.8	47	1.5	4.4
14	6.0	10.8	48	1.3	4.2
15	5.9	10.7	49	1.2	3.9
16	5.9	10.6	50	1.0	3.7
17	5.8	10.5	51	0.9	3.5
18	5.7	10.3	52	0.7	3.2
19	5.5	10.2	53	0.6	3.0
20	5.4	10.0	54	0.4	2.8
21	5.3	9.9	55	0.3	2.6
22	5.2	9.7	56	0.2	2.3
23	5.1	9.5	57	0.0	2.1
24	4.9	9.4	58	-	1.9
25	4.8	9.2	59	-	1.6
26	4.7	9.0	60	-	1.4
27	4.5	8.8	61	-	1.1
28	4.4	8.6	62	-	0.9
29	4.3	8.4	63	-	0.7
30	4.1	8.2	64	-	0.5
31	4.0	8.0	65	-	0.3
32	3.8	7.8	66	-	0.0
33	3.6	7.6	67	-	-

Studies that have cluster random assignment designs must meet the attrition standards for both the sample units that were assigned to the intervention or comparison group (for example, schools or communities) and the sample units for analysis (for example, youth attending those schools or living in those communities). In applying the attrition standards to the unit of analysis, the denominator for the attrition calculation includes only sample members in the clusters who remained in the study sample.

Criterion RCT.3: Was the probability of assignment into the research groups consistent over time, or if not, was the change accounted for? The rate at which sample members are assigned to the study's intervention and comparison groups does not have to be the same for both groups, but it must be consistent over time. For example, if a study began by assigning people to the intervention group 50 percent of the time, then increased that rate halfway through the study to 75 percent of the time, this could introduce bias into the estimated impacts if it is not appropriately controlled for. If there is no indication in the study that the authors varied the rates of random assignment over time, this criterion is satisfied. If they varied these rates, they must adjust for this in their analysis, which is most commonly done by applying weights to the analysis sample.

1. Note on imputation

If an RCT receives a high causal evidence rating, imputation of baseline or outcome variables is unlikely to result in biases to the estimated impacts. Thus, information about imputation will not be included in the study profile when studies are highly rated.

C. Interrupted time-series analyses

In ITS analyses, researchers examine the differences in outcomes over time, comparing the pre-intervention and post-intervention values of outcomes for a given unit. Rather than having separate intervention and comparison groups, each unit serves as a control for itself. The simplest and most common ITS analysis is a pre-post comparison, in which researchers compare outcomes observed at a single point in time before an intervention with those observed at a single point after an intervention.

A study should be reviewed under the ITS guidelines if (1) the comparison group contains the exact same units as the intervention group and (2) the two conditions simply represent different time points for measuring that group.[5] That is, the ITS criteria should only be applied to studies in which all units analyzed received the intervention and each unit serves as the *only* comparison for itself; CLEAR will evaluate studies using difference-in-differences and fixed effects analyses under the regression criteria (see Section A).

Although each unit of analysis serves as a control for itself in an ITS analysis, the unit may be composed of different individuals over time. For example, an analysis that compares the actions of new hires at a firm over time examines the same unit of analysis (the firm), even though the measure is composed of different individuals (the new workers).

[5] For example, although Langbein (2012) classifies difference-in-differences and fixed-effects analyses as ITS comparison group designs, CLEAR assesses those types of designs using the regression criteria because they include external comparison groups. Langbein, L. *Public Program Evaluation: A Statistical Guide.* Armonk, NY: M.E. Sharpe, Inc., 2012.

Like RCTs, ITS designs can receive high, moderate, or low ratings for causal evidence. A study gets a moderate rating if it meets the criteria for ITS.1, ITS.2a, and ITS.3; otherwise, the study receives a low rating. Studies also satisfying Criterion ITS.2b and Criterion ITS.4 can be highly rated.[6] Additionally, research designs that analyze groups of people over time (for example, states or companies) must satisfy Criterion ITS.5.

1. Criteria for all ITS designs

Criterion ITS.1: Was there no selection into the intervention based on pre-intervention trends in the outcomes of interest or the characteristics of participants? The study must provide evidence that pre-intervention trends in the outcomes or other observed or unobserved factors did not determine when the intervention was introduced. A study can show this in a variety of ways, including demonstrating that outcomes were stable before the intervention began. Alternatively, the study could account for any existing trends in its analysis. For example, suppose a company decided to begin auto-enrolling all new employees into its 401(k) program after observing recent declines in participation. Any resulting change in outcomes (plan enrollment) could simply be the by-product of the previous trend in participation (which was declining), so a study with this design would not meet this criterion.

Similarly, study authors must do one of two things to show that the intervention itself, and not the types of people who choose to participate in it, drives the results. First, an author could analyze the impact of the intervention on all people eligible to receive it, regardless of take-up. This would show how eligibility for the intervention, or intent-to-treat, affects the outcomes. For example, suppose an author analyzed the impact of a training program on wages. An analysis of the impact of program eligibility on all eligible people would meet ITS.1 (if the program was not introduced because of pre-intervention trends). Second, authors can analyze only the people who choose to participate in the intervention if the study shows that this group is similar to the group of all eligible people. That is, if authors only look at people who select into the intervention group, they must show that these individuals are comparable to other eligible participants on the observed characteristics specified in the relevant topic-area protocol.

Unobserved characteristics could also affect a person's choice to receive the intervention and influence the outcome of interest (endogenous selection mechanisms). For example, suppose that 100 people had the opportunity to participate in a job training program based on education and previous work experience, but only 50 enrolled in the program. Those who enrolled and those who did not might look similar on observed characteristics. But those who enrolled likely have some innate characteristic, such as motivation or ambition, that those who were eligible but chose not to enroll lack, and this trait would affect both participation in the program and post-intervention outcomes. If reviewers identify a plausible and documented selection mechanism that is not controlled for in the analysis, they indicate their concerns to the PI, who then seeks guidance from the CLEAR leadership team to determine whether the study meets Criterion ITS.1.

[6] Although ITS designs can receive high or moderate ratings for causal evidence, CLEAR leadership anticipates that only a small number of these studies in the topic areas examined by CLEAR will receive a moderate rating, and few (if any) studies will be highly rated. Therefore, any ITS study that receives a high or moderate rating during its first review will receive a second review by the topic-area PI or another senior reviewer.

Criterion ITS.2: Does the study include multiple demonstrations and multiple observations per demonstration? Following professional conventions (see Section E), CLEAR requires that the authors of an ITS analysis consider the effect of at least three distinct demonstrations of an intervention, in which a demonstration is defined as the introduction of the intervention to a unit of observation at a distinct time (Horner et al. 2012).[7] That is, authors must examine *at least three units of observation* that became subject to an intervention *at three different points in time*. This requirement is designed to limit the chance that changes in outcomes reflect some other factor that changed at the time of the intervention.

ITS analyses must also include multiple observations of each unit analyzed. Multiple observations from before the demonstration are required to ensure that pre-existing trends do not bias the results. Likewise, multiple observations of units after the demonstration are required to see whether the impacts were permanent or temporary and to determine the likelihood that the results represent a real change or are simply noise. The number of observations provided by the study affects the causal evidence rating.

- **Criterion ITS2.a: Threshold for a moderate causal evidence rating.** To be considered for a moderate rating, CLEAR requires that a study use data from *at least three points in time before and three points in time after the demonstration.*
- **Criterion ITS.2b: Threshold for a high causal evidence rating.** This criterion is similar to Criterion ITS.2a but requires study authors to use data from *at least five points in time before and five points in time after the demonstration.* (If a study does not meet Criterion ITS.2a, CLEAR does not assess ITS 2.b.)

Finally, the values observed before an intervention must be drawn from a sufficiently long period of time. This period varies based on the topic area and is specified in each review protocol. Periods will be chosen to ensure that existing trends in outcomes will not lead to incorrect conclusions. For example, Andersson et al. (2013) and Dyke et al. (2006) both showed that the earnings of workers receiving employment services tend to dip in the year before they enter the program, suggesting data more than a year before program participation would be needed.[8]

Criterion ITS.3: Was sample members' anticipation of the intervention either unlikely or appropriately controlled for? This criterion is identical to Criterion Regression.3. To satisfy this criterion, for example, an ITS study could document that the intervention happened at an unexpected time, or authors could explicitly note that subjects were not aware of a change in policy before the change was introduced.

---

[7] Horner, R., Swaminathan, H., Sugai, G., & Smolkowski, K. (2012). Expanding Analysis and Use of Single-Case Research. *Education and Treatment of Children,* 35, 269–290.

[8] Andersson et al. (2013) examined people receiving services through the Workforce Investment Act. See Andersson, F., Harry, H., Lane, J., Rosenblum, D., & Smith, J. (2013). Does Federally-Funded Job Training Work? Nonexperimental Estimates of WIA Training Impacts Using Longitudinal Data on Workers and Firms. NBER Working Paper No. 19446. Cambridge, MA: National Bureau of Economics Research.

Dyke et al. (2007) examined people receiving employment services in Missouri and North Carolina as part of Temporary Assistance for Needy Families. See Dyke, A., C. Heinrich, P. Mueser, K. Troske, and K. Jeon. "The Effects of Welfare-to-Work Program Activities on Labor Market Outcomes." *Journal of Labor Economics,* vol. 24, no. 3, 2007, pp. 567–607.

Criterion ITS.4: Was the intervention introduced at a predetermined time and in a predetermined manner? If a study does not meet Criterion ITS.1 through ITS.3, CLEAR does not assess the study for Criterion ITS.4; this criterion provides additional assurance that estimated intervention effects are not the result of selection. The intervention must be systematically introduced at a predetermined time and in a predetermined manner decided by the researcher.[9] The timing of the intervention need not be random to meet this criterion, but the study must show that the timing was deterministic and chosen by the researcher (and not by the entities examined by the study). For example, suppose a researcher was estimating the effect of a mentoring program on youth at risk of dropping out of school. If the organization providing the mentoring program selected when a youth became eligible for the program, the study would not meet Criterion ITS.4. If, instead, the researcher determined that the mentoring program would begin after the fifth week of the school year, the study would meet Criterion ITS.4.

Criteria ITS.1 and ITS.4 should be considered independently; satisfying one of the criteria does not imply satisfaction of the other. To assess whether a study meets ITS.1, a reviewer examines whether selection into the study sample, or selection of the timing of the intervention, could affect the study's findings. To assess whether a study meets ITS.4, a reviewer examines who determined when an intervention would be implemented. Consider again a study on the effect of a mentoring program on youth at risk of dropout. The study meets ITS.1 if participants in the program do not appear to differ from other eligible people and if the study authors account for any existing trends in a student's risk of dropping out that could trigger program enrollment (for example, by showing that a student's grades and school behavior did not change in the year preceding enrollment). The study meets ITS.4 if the authors determined the timing of the program (not the student, school, or organization providing the program).

2. Special criterion for ITS designs with group-level analyses

Criterion ITS.5: Were there no changes in group composition? This criterion is identical to Criterion Regression.4, but there are special considerations when applying it in an ITS context.

Reviewers will assess the selective migration in ITS studies based on both the people who join the group analyzed (joiners) and those who exit this group (leavers). Overall migration is defined as the number of joiners plus the number of leavers divided by the total number of people who were in the sample at any point (that is, joiners plus leavers plus those in the sample at both the beginning and end of the study period). Differential migration is the difference in the rate of migration from joiners and the rate of migration from leavers.

As mentioned under Criterion Regression.4, which addresses the same question, topic-area protocols can list group-level analyses that need not meet this criterion, with special consideration about how this situation would apply in an ITS setting. For example, consider a study on the effect of a certain company-level policy on workers' injuries. The policy mandates harsh sanctions on workers who fail to take appropriate safety precautions. The least safe workers might leave the company after the policy is implemented, implying changes in the company's composition. In this example, the compositional change could be considered part of the impact of the policy that might lead to changes in injury rates. Thus, the protocol would waive Criterion ITS.5 for company-level analyses of this type.

[9] This criterion can be thought of as similar to the requirement that the running variable may not be manipulated by an individual in a regression discontinuity design. See McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics 142*(2), 698–714.

D. Considerations for all profile reviews of causal studies

A study author must make a convincing case that each criterion is satisfied; reviewers check for plausible scenarios documented in the study under which a criterion is not satisfied. But because there is no definitive test to indicate that some of the criteria have been met, CLEAR leadership must occasionally help reviewers make a determination that is consistent across CLEAR, and consistent with the spirit and intent of the evidence guidelines. When questions arise as to whether a study meets a given criterion, reviewers submit their concerns to the topic area's PI. The PI recommends an appropriate interpretation of the CLEAR guidelines, and this interpretation is reviewed and confirmed (or modified) by the CLEAR leadership team, as described in *CLEAR Policies and Procedures,* version 3.0. To promote consistency and transparency, these decisions are documented and become the basis for future refinements and clarifications to the CLEAR evidence guidelines (or, if specific to a topic area, documented in the final protocol for the topic area).

Several other considerations affect the interpretation of a study's findings but do not affect the causal evidence rating:

- For analyses involving instrumental variables, the reviewers consider whether standard errors were calculated using an appropriate method, such as the delta method or bootstrapping, and whether, if necessary, standard errors accounted for the first stage of estimation in a two-step process.
- For analyses involving longitudinal data, the reviewers consider whether the authors calculated standard errors using a method that accounts for serial correlation, heteroskedasticity, and different levels of aggregation (for example, cluster-robust standard errors for designs that use individual and state data).
- For all designs, the reviewers consider whether the authors estimated multiple related impacts, which makes it more likely they will find some statistically significant differences simply by chance, and whether the imputation of baseline variables or outcome variables could cloud the interpretation of a study's findings.

If the authors did not address these issues appropriately, the study profile produced by CLEAR notes this as a concern. Profiles might also note when a contextual factor or implementation issue might affect the interpretation of the study's findings and mention what the authors could have done to receive a higher causal evidence rating.

E. Developing and clarifying CLEAR guidelines for causal evidence

In collaboration with the Department of Labor (DOL) and a technical work group (TWG) of experts, during the first phase of the project CLEAR developed a set of guidelines for reviewing non-experimental research with causal designs. These guidelines focus on various regression analyses, including those with fixed or random effects as well as difference-in-differences and instrumental variables. During CLEAR's pilot phase, the evidence guidelines underwent a continual review and improvement process and were revised to reflect lessons learned as CLEAR first implemented the guidelines. Version 1.1 incorporated these revisions, along with feedback from DOL and the TWG.

CLEAR uses the WWC evidence standards, adapted for use in a labor context, to evaluate the strength of causal evidence of RCTs. These standards have been extensively reviewed and tested, and they represent the current state-of-the-art in rating the strength of RCTs.[10]

During CLEAR's second phase, the clearinghouse developed evidence guidelines for evaluating the quality of causal evidence for studies with ITS designs; version 2.0 incorporated these guidelines.[11] They draw from the WWC evidence standards for single-case designs (SCDs).[12] Experts in their fields developed the WWC standards, which are also extensively informed by the literature on how to best conduct analyses of SCDs. For example, SCDs rely upon replication to provide causal evidence. Typically, the larger the number of units examined in a given study, the more confidence one can have in the study's results (Kratochwill and Levin 2010).[13] Similarly, CLEAR Criterion ITS.2 requires multiple units of study. In addition, CLEAR consulted with experts on both SCDs and ITS designs to ensure that the adaption of WWC criteria was appropriate.

The criteria in these guidelines are used to review evidence from research papers and reports that span a broad range of social science disciplines over many years. For this reason, with a few exceptions, the guidelines do not require specific approaches. Rather, they provide general criteria that must be met; a supporting addendum provides examples (gleaned from completed CLEAR reviews) showing how the criteria could be satisfied.

Version 2.1 includes minor refinements to the Version 2.0 standards, such as clarifying the process CLEAR uses to resolve unique or challenging methodological issues. It also reorders some criteria for ease of understanding.

---

[10] The full set of WWC evidence standards is documented in the *WWC Procedures and Standards Handbook*, version 3.0, which is available at http://ies.ed.gov/ncee/wwc/documentsum.aspx?sid=19. The handbook explains the criteria for evaluating RCTs, which mainly involve determining study attrition and any other threats to the validity of the study's design. Other federal research clearinghouses have adapted the WWC standards, including the U.S. Department of Health and Human Services (HHS) for the Teen Pregnancy Prevention evidence reviews, the Institute of Education Sciences for the evaluation of Investing in Innovation (i3) evidence, and the HHS Office of the Administration for Children and Families for the Home-Visiting Evidence of Effectiveness systematic reviews.

[11] Regression discontinuity designs are not included in the causal evidence guidelines because there has not been a need for guidelines for these designs thus far. CLEAR will develop such guidelines in future phases of the project as necessary.

[12] These criteria were also informed, to a lesser extent, by the WWC's standards for regression discontinuity designs, developed in a parallel manner to those for SCDs. Those standards have been piloted since 2010 and are currently detailed in the *WWC Procedures and Standards Handbook,* version 3.0 (available at http://ies.ed.gov/ncee/wwc/documentsum.aspx?sid=19).

[13] Kratochwill, T., and J. Levin. "Enhancing the Scientific Credibility of Single-Case Intervention Research: Randomization to the Rescue." *Psychological Methods,* vol. 15, 2010, pp. 124–144.