ADDENDUM TO THE CLEAR CAUSAL EVIDENCE GUIDELINES, VERSION 2.0

As noted in the Clearinghouse for Labor Evaluation and Research (CLEAR) Evidence Guidelines, version 2.0, determining whether a given criterion has been met sometimes requires judgment on the part of a reviewer because there is no definitive test to indicate that it has been met. The guidelines specify that the study must make a convincing case that the criterion is satisfied and, if reviewers can identify a plausible scenario under which it is not, the study fails this test (subject to the review of the principal investigator [PI]). This addendum provides additional guidance on satisfying the CLEAR criteria for causal designs, in particular for nonexperimental designs. It provides suggested tests, when applicable, and examples to illustrate the concepts.

A. Instrumental Variables Models

Satisfying Criterion IV.1: Sufficient Instrument Strength (or Instrument Relevance)

Studies must report a test of instrument strength to satisfy this criterion. A common test uses a first-stage equation that models receiving the 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 receiving the intervention. The F-test must be conducted only for the instrumental variables in the first stage; other covariates must be excluded from the F-test, because their relationship with the outcome is irrelevant to the question of instrument strength. Stock and Yogo (2005) provided evidence that the F-statistic should exceed 16.38 to consider the instrument to have sufficient strength. However, to be consistent with labor studies from the 1990s that use instrumental variables techniques, CLEAR requires that the F-statistic exceed 10.00 for the instrument to be considered sufficiently strong, if this approach is used.

Other approaches include testing the null hypothesis that the forward and reverse two stage least squares estimates are the same (Hahn and Hausman, 2002) and the Anderson canonical correlation analysis (Anderson, 1984).

Note that older studies (those published before Bound et al., 1995) might not report tests of instrument strength. Thus, for studies published in 1995 or earlier, reviewer discretion (subject to the review of the PI) will be used to determine whether the instrument is likely of sufficient strength.

Satisfying Criterion IV.2: Exclusion Restriction (or Instrument Exogeneity)

Several tests of the exclusion restriction are informative; however, none provide conclusive evidence that the restriction is valid. Examples include the Wu-Hausman test and the Sargan test (also known as the Hansen test; see Greene, 2011). Another approach to testing the exclusion restriction is to regress the instrument on exogenous variables. Angrist and Krueger (1999) provided a discussion of falsification tests (also known as placebo or refutability tests).

In general, instrumental variables that are determined by random factors, such as lotteries, satisfy the exclusion restriction. For example, suppose a study estimated the effect of participating in a job training program on future wages. The problem is that participation in the program might be endogenous, as people with higher intrinsic motivation might both be more likely to participate in the program and have higher future wages. But suppose that the job training program was oversubscribed and a lottery determined admission. In this case, a variable indicating whether the individual "won" admission to the program satisfies the exclusion restriction; the only way winning

the lottery could affect future wages is through the individual being able to participate in the job training program.

Cases that do not use a random factor as an instrument are less clear. As noted in the guidelines, studies do not satisfy the exclusion restriction if the reviewer identifies a plausible mechanism by which the instrument would influence the outcome directly (after controlling for observed factors). For instance, suppose a study estimated whether workplace inspections conducted by the Occupational Safety and Health Administration (OSHA) reduced the rate of workplace injuries. A simple regression with the rate of workplace injuries as the dependent variable and an indicator variable for whether the firm was inspected as the key explanatory variable of interest would be biased because inspections are not always done at random; firms that have reported a workplace injury are more likely to be inspected. Therefore, authors might seek to use an instrumental variables approach to account for the endogenous inspection variable. Suppose the authors used an indicator for whether the workplace had been inspected by the Environmental Protection Agency (EPA), which could predict OSHA inspections if, for instance, there were an overall initiative to prioritize enforcement actions across all government agencies. Although the presence of an EPA inspection seems unrelated to workplace injuries at first glance, it might capture unobserved factors such as poor company management or a general disregard for regulations that could trigger both kinds of inspections and have an effect on workplace injuries. Therefore, this instrument does not satisfy the exclusion restriction.

Satisfying Criterion IV.3: Identification

Studies with multiple endogenous variables and instruments may test whether the rank condition holds by running reduced-form regressions of the dependent variable on the exogenous variables (including the instruments) and performing an F-test or Wald test of the null hypothesis that all of the coefficients on the instrumental variables are equal to zero; this is necessary, although not sufficient, for the rank condition to hold. Alternatively, the study could test the hypothesis that the rank is of a given size against the alternative that the rank is larger, as proposed by Cragg and Donald (1996).

Not all studies can explicitly test the rank condition. If this test is not provided, reviewers will compare the number of instruments representing unique variables with the number of endogenous variables. If this number of instruments is greater than or equal to the number of endogenous variables, then this criterion is satisfied.

Following the previous example, suppose an author is interested in whether OSHA inspections reduce the rate of workplace injuries. If the endogenous independent variable of interest is an indicator for any OSHA inspection and the authors used EPA inspections as an instrument, the identification criterion is satisfied; there is one endogenous variable and one instrument derived from one unique variable. However, suppose the authors were interested in two endogenous variables: any OSHA inspection and type of most recent OSHA inspection. If they used number of EPA inspections and an indicator for any EPA inspection as instruments, this would not satisfy Criterion IV.3 because the two instruments are simply different representations of the same underlying variable.

Other Considerations

Three additional factors will be considered in the instrumental variables study review process: (1) correctly estimating standard errors, (2) showing the appropriateness of the instrumental

variables approach, and (3) estimation strategy. However, these are not criteria; rather, if there are concerns about these issues, they will be noted as cautions in the study summary.

1. The standard errors of instrumental variables estimates obtained by using a two-step estimation procedure are typically incorrect, and smaller than the correct ones, unless the variance associated with using the predicted values from the first stage is taken into account. Thus, standard errors for these types of estimates should use an estimate of the covariance matrix that takes account of the first stage. In addition, if the instrument is weak, the two-stage least squares estimate of the variance can perform poorly.
2. Studies using instrumental variables should also present the results without instrumenting for endogenous variables to show what is gained by using an instrumental variables approach. This can corroborate the appropriateness of using the method.
3. Instrumental variables models can be estimated using two-stage least squares methods or generalized methods of moments. Studies should include a justification for the method used.

B. Regression Analyses

Satisfying Criterion Regression.1: Comparability of Treatment and Control Groups Before the Intervention

To attribute any differences in outcomes between the treatment and control groups to the intervention, it must be the case that the treatment and control groups would have had similar outcomes in the absence of the intervention. This criterion tests the reasonability of this expectation. In cross-sectional regressions, whether this criterion is met is determined by assessing whether the treatment and control groups had similar observable characteristics before the intervention. The topic area protocols describe on which characteristics the groups must be similar. For studies using panel data, including difference-in-differences models and those with individual fixed effects, whether this criterion is met is determined by assessing whether the treatment and control groups experienced similar trends in outcomes before the intervention.

Observed characteristics. To establish comparability of the groups on observed characteristics in a cross-sectional regression, study authors could compare characteristics measured before the intervention for the two groups and show that the differences between the two are not statistically significant. For instance, they could do the following:

1. Use chi-squared or t-statistics to demonstrate that standardized mean differences between treatment and control groups (the difference between the group means divided by the pooled within-group standard deviation) are not statistically significant ($p > 0.05$).
2. Demonstrate that means for key variables are not statistically significantly different ($p > 0.05$) across treatment and control groups.

If the authors do not attempt this type of demonstration, or if the groups do appear to be different, then study authors must control for these characteristics in the analysis. Typically, basic demographic information alone will not suffice to establish comparability of the groups or as sufficient controls in a regression; pre-intervention and/or lagged values of the key outcome measure will usually be necessary to satisfy this criterion. The number of lags in the pre-intervention outcomes and the types of control variables required vary by the topic area and outcome being examined. For example, for studies that analyze the employment outcomes of youth under the

Opportunities for Youth protocol, the specified pre-intervention characteristics include pre-intervention measures of employment (lagged employment variables), age, gender, race/ethnicity, and geographic location. If pre-intervention measures of the outcomes are not available, judgment is required as to which control variables, and the appropriateness of the functional form they take, are sufficient. In addition, if pre-intervention differences in key variables are large, controlling for those variables might not be sufficient to remove bias in the estimation. Reviewers and the PI will consider these situations on a case-by-case basis.

For studies using panel data, including difference-in-differences models and those with individual fixed effects, authors must demonstrate equivalent *trends* between treatment and control groups to satisfy this criterion. This applies both to a single policy change affecting all parties at once (for instance, a change in a law or regulation that is applied to an entire region or set of industries) with multiple pre-intervention time periods and to the case of different parties choosing to adopt a policy at different times (for example, states adopting minimum wages throughout the early 2000s). Authors can demonstrate common or parallel trends in the pre-intervention measures of outcomes by doing the following:

1. Plot the trend lines. Reviewers will examine the plotted trend lines visually to determine whether this criterion is satisfied. If the slopes of the trend lines in the pre-intervention period differ enough that they are visually discernible, and the authors did not adjust for these differences using control variables, the study will not be considered to meet this criterion.

2. Use descriptive statistics, such as computing the change in outcomes over time within a regression framework, to demonstrate there are no statistically significant ($p > 0.05$) differences between the slope coefficients for the treatment and control groups in the pre-intervention period.

Consider an analysis of panel data examining the impact of programmed OSHA inspections on injury rates. To satisfy this criterion, a study comparing inspected and uninspected firms and including firm-level fixed effects in the regression would have to demonstrate that, before the inspections began in the treatment group, injury rates in the firms in the treatment and comparison groups had evolved in a similar way over time.

If the authors analyzed only two time periods—one before and one after the policy change—it would not be possible to study the pre-treatment trend in outcome variables. Therefore, studies that use only one pre-intervention time period can satisfy this criterion by performing some type of placebo test to indicate that the comparison group is an appropriate counterfactual. For example, suppose a study estimated the effect of a 2010 change in the minimum wage law in one state on the unemployment rate and used a neighboring state that did not change the minimum wage law in 2010 as the comparison. The authors could examine the change in the unemployment rate from the year before the law was enacted (2009) to the year after (2010) for both states. In this case, they could perform a placebo test in which they computed the estimate on the unemployment rate in particular industries that would not have been affected by the change in minimum wage laws, such as skilled manufacturing.

If the authors do not attempt a demonstration of equivalent trends, or do attempt a demonstration but show differences in trends, then they must adequately control for time-varying characteristics that might influence the outcomes of interest. The specific controls depend on the

model being estimated, but could include a time trend, time fixed effects, or individual-specific time trends and a series of time fixed effects.

Unobserved characteristics. Another consideration for studies to meet this criterion is comparability of the treatment and control groups on unobserved characteristics. This guards against the possibility that the treatment and control groups appear to be similar on observed characteristics, but there is an obvious selection mechanism whereby people select to enter the treatment group based on a characteristic that is not observed and influences both the decision to participate and the outcome of interest. For example, suppose applicants who meet basic screening criteria in terms of education and previous work experience receive offers to participate in a job training program. Then, only half of those offered participation in the program actually complete the program, with the other half dropping out. In this instance, those who completed the program and those who dropped out might look very similar on observed characteristics. However, those who actually completed the program likely have some innate characteristic, such as motivation or ambition, that those who dropped out do not possess, and this trait would influence both participation in the program and post-intervention outcomes. Therefore, using the dropouts as a comparison group for those who completed the program does not provide a valid picture of how the program completers would have fared in the absence of the offer to participate in the program. A study using this approach would not satisfy this criterion.

In general, any intervention triggered by changes in the outcome variable will likely have issues related to noncomparability of unobservable characteristics. For example, consider an analysis of the impact of receiving an OSHA citation on the number of subsequent injuries at a firm. OSHA citations are a function of the number and type of hazards at a firm, which in turn influence injury rates. Thus, it is unlikely that a study of differences between firms receiving and not receiving OSHA citations could meet Criterion Regression.1.

For fixed effects models, satisfying this criterion requires that there should be no other obvious selection mechanism into the intervention when the model includes the fixed effects—which control for unobserved, time-invariant factors—and key time-varying factors. In other words, there should be no other unobserved, time-varying factors that drive selection into the intervention *and* influence the outcomes of interest. For example, OSHA inspections can be triggered by complaints of hazards, which vary at a firm over time. But the presence of hazards also influences the injury rates at the firm. Thus, a model analyzing OSHA inspections must include controls for changes in hazards over time (which are potentially unobservable to the researcher). Alternatively, consider a job training program that is available only to spouses of deployed military personnel. To meet Criterion Regression.1, the model must include controls for the spouse's deployment status, which changes over time and may also affect the nondeployed spouse's earnings by, for example, increasing the nondeployed spouse's child care responsibilities. Thus, fixed effects can be used to control for only those factors that are shown or known to vary little over time.

To deal effectively with self-selection, a model could do one of the following:

1. Demonstrate that it is not an issue, for example by regressing outcome residuals on indicator variables for time periods before program participation.[1] It could also be accomplished using a graphical or descriptive analysis.
2. Adjust for it appropriately using a flexible set of control variables. These adjustments could be made by including key independent variables in the regression or using a propensity-score approach, in which the propensity score is calculated using these key variables.

Note on matching designs. CLEAR evaluates matching designs, including propensity-score matching designs, according to the guidelines for regression analyses. These studies must match on all the control variables specified in the topic area protocol. Studies that match on all the specified control variables do not also have to include them in the regression as long as the match was successful. If the study does not match on one or more key control variables, or if the matching process was unsuccessful on one or more key control variables, the study must include them in the regression analysis. A matching process is determined to be unsuccessful if, after matching, there remain statistically significant differences between the treatment and control groups at the 5 percent level.[2]

Satisfying Regression.2: Confounding Treatments

This criterion can be difficult to judge and reviewers must use their knowledge of the intervention being examined by the study, the political environment, and any other background in assessing whether the model does not account for any plausible confounding treatments. There are some cases in which it is fairly obvious that a confounding or omitted treatment exists. For instance, suppose a study estimated the effect of state minimum wage laws on the unemployment rate and examined this using a model with state-level policy variation. Suppose that states that adopted minimum wage laws simultaneously adopted laws relating to mandatory employer provision of health care, which might also have an effect on the unemployment rate. This would confound estimates of the causal impact of the minimum wage law. Note, however, that if only some of the states simultaneously enacted other policies related to the outcome, the causal interpretation is not necessarily invalid. In addition, if the two treatments can be considered a bundled treatment—in which case the bundle would be replicable—this does not necessarily represent a confounding treatment. In practice, confounds in the labor literature are relatively rare, and reviewers should raise potential confounds with the topic area PIs to make a final determination.

---

[1] Such a regression would yield a negative coefficient on the time period before enrolling in the program if a negative earnings shock is present. However, if the coefficients on these variables are indistinguishable from zero, past shocks to the outcome are uncorrelated with program participation.

[2] Although it is common in research to consider effect size magnitude to evaluate the success of a matching approach, CLEAR strives to be consistent in the criteria for matching and other types of regression designs. CLEAR leadership anticipates that many studies will provide neither effect sizes nor the information that CLEAR reviewers would need to compute them. Furthermore, for consistency with other designs, CLEAR criteria do not address other common practices in strong matching designs, such as correctness of the variance estimator. The CLEAR criteria for using statistical significance as a basis for evaluating the success of a matching approach is consistent with other federal evidence reviews, including the Department of Health and Human Services' Pregnancy Prevention Research Evidence Review and Home Visiting Evidence of Effectiveness Review.

Satisfying Regression.3: Anticipating the Intervention

Under some circumstances, knowledge that a policy will be enacted or that one will be able to participate in a program can alter behavior in ways that can threaten the validity of a study. For instance, suppose that participants who qualified for a job training program reduced their effort at work in the months preceding the start of the training program, knowing that they would be leaving the job soon. Studies that used pre-intervention data covering the period of reduced effort would incorrectly estimate the impacts of those training programs on future wages because participants were starting at lower earnings levels than they would have in the absence of anticipating the training program. As another example, suppose a new state safety standard was announced that would go into effect in six months. During the six months between policy adoption and enforcement, businesses might begin increasing their safety precautions in anticipation of the new law, potentially reducing their rate of workplace injuries; a study that used six-month lagged data as controls in the regression model could therefore yield biased impact estimates.

In some cases, anticipation is likely not a threat to identification. However, if the reviewer identifies a plausible mechanism by which anticipation could influence the study's findings, study authors must do one of the following:

1. Demonstrate that it is not an issue. In the example of adopting a new safety standard, this could be accomplished by reporting precise dates of when the standard was announced, inspections were conducted, implementation began, and pre-intervention data were collected. It could also be accomplished using a graphical or descriptive analysis.
2. Adjust for it appropriately—for instance, by removing time periods that exhibit treatment anticipation from the analysis.

Satisfying Criterion Regression.4: Changes in Group Composition

Studies that include group-level rather than individual-level fixed effects (for instance, using data from the Current Population Survey on people's earnings to evaluate the effect of state minimum wage laws) or group-level control variables to account for pre-intervention lags in the outcome variable must provide evidence that there was not substantial selective migration into or out of the study sample. CLEAR uses the What Works Clearinghouse (WWC) attrition standards to determine whether the overall and differential changes in group composition are a potential source of bias in the impact estimates. If overall and differential changes in composition are shown to be within the cutoffs established by the WWC for attrition standards, the study does not have to make additional adjustments. However, if they exceed the cutoffs, data on individual or group characteristics must be used to account for measurable changes in composition.

In cases in which group composition can reflect the impact of an intervention (specified in the topic area protocol), this criterion might not apply. That is, changes in group composition due to the intervention can be thought of as an alternative mechanism for an intervention to influence outcomes. For example, in analyses of OSHA enforcement activities using state- or industry-level data, Criterion Regression.4 does not apply. Any changes in the composition or characteristics of firms in the aggregate data following an OSHA enforcement activity can be seen as an impact of that activity and thus should be part of the measured treatment effect. Therefore, studies in this area need not demonstrate that the treatment left the group composition unchanged.

Satisfying Criterion RE.1: Justify the Use of Random Effects

To meet this criterion, studies must include a test justifying the use of random effects over fixed effects. This is because fixed effects estimates will be unbiased even if the assumptions underlying use of random effects are violated (Johnston and DiNardo, 1997). One common test is the Wu-Hausman test, in which the model is estimated using both random and fixed effects, and the estimated coefficients and covariance matrices from both are used to generate a test statistic used to test the null hypothesis that the random effects model is correct (see Greene [2011] for details).

Other Considerations

Three additional factors will be considered in the regression study review process: (1) estimating standard errors, (2) accounting for the nature of the dependent variable (for example, binary dependent variables), and (3) imputation of baseline and/or outcome variables. However, these are not criteria; rather, if there are concerns about these issues, the study summary will note them as cautions.

1. The reviewers will consider whether standard errors were calculated in a manner that addressed differential levels of aggregation, heteroskedasticity, or serial correlation, if relevant. Regarding standard error corrections, Donald and Lang (2007) showed that not accounting for heteroskedasticity in panel data models also understates the standard deviation of the estimates. Bertrand et al. (2004) showed that difference-in-differences models that do not address serial correlation severely understate the standard deviation of the estimators and provided evidence on corrections for this problem. Examples of methods of addressing these issues include bootstrapping and computing standard errors robust to serial correlation and heteroskedasticity.
2. The reviewers will consider whether the specification properly accounts for the nature of the dependent variable. For example, when the outcome is a binary dependent variable, nonlinear models such as probit or logit models might be more appropriate than ordinary least squares.
3. The reviewers will discern whether the authors imputed baseline and/or outcome data for some analysis sample members. Doing so can introduce biases into the estimated impacts and/or the interpretation of the study's findings.

In addition, reviewers can note other issues about the study that they believe readers of the study should know, such as diluted treatment effects, nonresponse bias, alternate explanations of the results, or whether the direction of the potential biases supports the authors' conclusions. For example, suppose a study finds that establishments inspected by OSHA did not have lower injury rates than establishments not inspected and interprets this finding as evidence that OSHA is ineffective. An alternative explanation is that OSHA could have effects on the safety level of all firms (general deterrence), as opposed to only the firms that receive inspections (specific deterrence). Reviewers could note this alternative explanation in this section.

C. Randomized Controlled Trials

Satisfying Criterion RCT.1: Sample attrition. Reviewers use an Excel spreadsheet that has been programmed with the attrition formulas to compute sample attrition at the units of assignment and analysis for a given attrition standard (conservative or liberal), which is specified in the topic area protocol. The baseline sample sizes are the numbers of people that were randomly assigned to

each study condition. The analysis sample sizes are the numbers of people that were included in the analysis of impacts.

In most cases, any differences between baseline and analysis sample sizes are counted toward attrition. This includes cases in which study participants were first assigned to study groups and then consent was sought; non-consenters count as attrition because their decision whether to consent to the study could have been influenced by the study group to which they had been assigned. It also includes instances of missing outcome data that arise for any reason, including survey nonresponse or missing administrative data. Thus, analysis sample sizes can vary for outcomes within the same study because of missing data on those outcomes or the use of different data sources. Correspondingly, some outcomes might be assessed as having high attrition whereas others within the same study might be assessed as having low attrition; in such cases, the study can receive two causal evidence ratings and the study summary indicates for which outcomes each rating applies.

Reviewers should not assume that use of administrative data implies complete coverage of the study participants and, thus, no attrition. Rather, they should find information in the study indicating how missing administrative variables were treated (for instance, assuming a zero for all missing values) or submit a query to the study authors to clarify.

Satisfying Criterion RCT.2: Confounding factors. See discussion for satisfying Regression.2.

Satisfying Criterion RCT.3: Consistent probabilities of assignment into the research groups. The rate at which individuals are assigned to the study's treatment and control 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 treatment 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 it appears that the probability of assignment into the research groups varied over the study period, authors must explicitly discuss what methods they used to account for the varying probabilities. Methods considered sufficient include using weights that reflect the overall probability of assignment to the treatment group and including controls for the times during which random assignment probabilities differed. If they did not mention any methods, CLEAR will submit an author query to clarify.

D. Interrupted Time Series Analyses

Satisfying Criterion ITS.1: Selection Into the Study and Intervention

To satisfy Criterion ITS.1, a study must demonstrate that neither selection into the group of people receiving the treatment nor the timing of the intervention (selection into the intervention) threatens the interpretation of the evidence. These different selection-related issues are typically assessed separately, though they lead to similar biases.

1. Selection Into the Study Sample

The authors should demonstrate that a person's choice to receive or complete the treatment does not drive the results of their analysis. Selection into the study sample occurs if the authors do not analyze data for all people eligible to receive a treatment. For example, an author analyzing a training program might focus only on the set of those completing the training program and not all

people who were eligible to enroll in the program. This could lead to biased results if program completers differed from dropouts in ways that were related to the outcomes examined.

Selection into the study sample should be assessed in a manner parallel to selection into the treatment under Criterion Regression.1, comparing the units of observation included in the study to those excluded. As in the regression criterion, the topic area protocol will provide a list of the observable characteristics on which the groups must be similar (these are identical to the characteristics required by Regression.1).

Observed characteristics. To establish comparability of the groups on observed characteristics, study authors could compare characteristics measured before the intervention for the two groups and show that the differences between the groups are not statistically significant. For instance, they could do the following:

1. Demonstrate that standardized mean differences between the study sample and other eligible units of observation (the difference between the group means divided by the pooled standard deviation) are relatively small (less than 0.25).
2. Demonstrate that means for key variables are not statistically significantly different ($p < 0.05$) across groups.

For example, consider an analysis of the impact of OSHA inspections on injury rates using a subset of firms that volunteered to receive an inspection. To demonstrate that analyzed and nonanalyzed firms (that is, firms that volunteered to receive an inspection and those that did not) are equivalent, a study would have to demonstrate that before the inspections began, the firms in both groups had injury rates that were similar. Additionally, the study would have to show the equivalence of the firms based on geographic location, firm size, and sector.

Unobserved characteristics. Another consideration for studies to meet this criterion is comparability of the analyzed group with other people eligible to receive the treatment based on unobserved characteristics. This guards against the possibility that the analyzed group appears similar to the eligible population on observable characteristics, but there is an obvious selection mechanism whereby individuals become part of the study sample. For example, suppose that offers to participate in a job training program are given to applicants who meet basic screening criteria in terms of education and previous work experience. Then, only half of those offered participation in the program actually complete it, with the other half dropping out. In this instance, a study analyzing only program completers would potentially reveal unrepresentative impacts. A study could instead analyze changes in outcomes for all those offered the chance to enroll in the program.

Note that CLEAR would not consider sample selection a problem if authors analyzed a treatment targeted to and received by a specific group of people. For example, if a study reviewed under the Opportunities for Youth protocol considered the impact of becoming eligible to enroll in a program that targeted at-risk Hispanic youth, the targeting of the program toward Hispanic youth and the omission from the analysis of youth from other ethnic backgrounds would not be a concern. However, if the study analyzed only a subsample of the at-risk Hispanic youth offered the chance to participate in the program, this would be cause for concern. Additionally, selection into the sample should not be classified as an issue if a well-defined subgroup of people is of interest for the study. For instance, in the previous example, a follow-up study might examine outcomes only for women offered the chance to participate in the program. Although this selects a subset of the eligible

population, it can produce an unbiased estimate of the impact of the program for women. Thus, this case would not violate ITS.1.

2. Selection Into the Treatment Group at a Specific Point in Time

To satisfy ITS.1, studies also have to demonstrate that selection into the intervention does not lead to biased results. The classic example of this issue comes from Ashenfelter (1978). This study first documented that people typically enter training programs after a period during which they have experienced particularly poor employment outcomes. Because of this, even in the absence of a training program, outcomes would probably improve over time for the population considered.

A study may demonstrate that the timing of an intervention does not bias the study's results in a variety of ways:

1. Demonstrating that trends in outcomes were flat before the initiation of the intervention
2. Arguing that institutional factors not related to levels or trends in outcomes, or other important factors, determine the timing of the intervention
3. Demonstrating that pre-intervention outcomes and other observable characteristics cannot predict the timing of the intervention

For example, consider a study that analyzes an after-school program encouraging girls in high school to explore careers in science, technology, engineering, and math (STEM) fields. To demonstrate that the timing of the intervention was not an issue, the study might argue that the schools that implemented the program all began it in September (with the beginning of the school year), that this timing was determined by the government agency funding the program, and that the year the implementation began was chosen without reference to the schools' preferences or profiles. Alternatively, a study that analyzed the impact of this program on the share of females enrolling in advanced coursework in mathematics or science might show that this proportion was roughly constant before the implementation of the after-school program. Finally, the study could also run a regression predicting the timing of the program's implementation based on school characteristics and prior outcome values. If these characteristics do not predict when the intervention occurred, ITS.1 is not violated.

In all cases, the reviewer should also use his or her discretion in determining if unobservable variables could determine the timing of an intervention. If there is some likely underlying cause for an intervention that is not demonstrated to be constant before implementation, a study likely does not meet criteria ITS.1.

Satisfying Criterion ITS.2a: Multiple Demonstrations and Multiple Observations Per Demonstration

The simplest ITS analyses use data on a very small number of people or firms. Although ITS designs can typically use fewer units of observation than other studies to provide causal evidence, a minimum amount of data is still required. In particular, to satisfy ITS.2a, a study must include three demonstrations, with a demonstration defined as the introduction of the intervention to one or more units of observation at a distinct time. Further, to meet the requirements of ITS.2a, a demonstration must also be analyzed using at least three pre-intervention observations and at least three post-intervention observations. Although a demonstration can introduce the intervention to only one individual, demonstrations involving multiple people cannot be counted multiple times.

That is, the introduction of an intervention at time *t* counts as a single demonstration, regardless of how many units of observation began receiving the intervention at time *t*. Additionally, the topic area protocol will specify the minimum amount of time that the pre-treatment interventions must cover. For example, a topic area examining employment and training programs could require ITS.2a to be demonstrated based on at least six quarters of pre-intervention data, to avoid issues identifying potential variants of Ashenfelter's dip in the analysis of training programs.

Satisfying Criterion ITS.2b: Multiple Demonstrations and Multiple Observations Per Demonstration

ITS.2b will be assessed using the same type of analysis as ITS.2a; however, demonstrations must include five pre-treatment and five post-treatment observations to satisfy ITS.2b. Three demonstrations are still required to satisfy ITS.2b and the requisite time period spanned by the pre-treatment data is specified by each topic area protocol.

Satisfying Criterion ITS.3: Anticipating the Intervention

This criterion is assessed in much the same way as Regression.3 and is designed to avoid similar issues. Under some circumstances, knowledge that a policy will be enacted or that one will be able to participate in a program can alter behavior in ways that can threaten the validity of a study. For instance, suppose a new state safety standard was announced that would go into effect in six months. During the six months between policy adoption and enforcement, businesses might begin increasing their safety precautions in anticipation of the new law, potentially reducing their rate of workplace injuries; a study that analyzed data starting only three months before the change might produce misleading results.

In some cases, anticipation is likely not a threat to identification and reviewer discretion should be used in assessing ITS.3. However, if the reviewer identifies a plausible mechanism by which anticipation could influence the study's findings, study authors must do one of the following:

1. Demonstrate that it is not an issue. In the example of adopting a new safety standard, this could be accomplished by reporting precise dates of when the standard was announced, inspections were conducted, implementation began, and pre-intervention data were collected. It could also be accomplished using a graphical or descriptive analysis.
2. Adjust for it appropriately—for instance, by analyzing the announcement of the future change in policy, rather than the implementation of the policy itself.

Satisfying Criterion ITS.4: Changes in Group Composition

Similar to Regression.4, ITS studies that analyze group-level data (for example, firms, schools, or cohorts) must demonstrate that changes in the composition of the groups analyzed are not responsible for the results. To assess ITS.4, CLEAR uses the WWC attrition standards but applies these to two different kinds of migration: people who join the group analyzed (joiners) and people who leave the group analyzed (leavers). Overall migration is defined as the number of joiners plus the number of leavers divided by the total number of people in the group at either the beginning or end of the observation period. Differential migration is the difference in the rate of migration from joiners (the share of the final analysis sample that joined after the intervention occurred) and the rate of migration from leavers (the share of the baseline sample that left after the intervention occurred).

If overall and differential migration exceed the WWC cutoffs for overall and differential attrition, a study must account for these changes to satisfy ITS.4.

As for Criterion Regression.4, topic area protocols may provide a listing of group-level analyses that need not meet this criterion. For example, consider a study that examines the effect on injuries of a firm-level policy that implemented harsh sanctions on workers who fail to take appropriate safety precautions. The least-safe workers might leave a firm after the policy is implemented, implying changes in the firms' 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.4 for firm-level analyses of this type.

Satisfying Criterion ITS.5: Manipulation of Treatment Status

This criterion, not required to rate a study as providing moderate causal evidence, assesses whether the timing of an intervention was determined by the researcher. To meet ITS.5, a study must examine an intervention that was introduced at a time and in a manner that was predetermined by the research team. Further, neither implementers nor the people or units analyzed can manipulate when the intervention is introduced. Alternatively, researchers could specify a protocol for implementers to use in determining when an intervention should be implemented. So long as implementers faithfully follow the criterion, the study can meet ITS.5.

For example, consider an analysis of the impact of 401(k) auto-enrollment policies on rates of participation in 401(k) plans. A study in which a firm chooses when to implement the auto-enrollment policy would not meet Criterion ITS.5. A study in which the researchers determined when the policy would be implemented at the firm would meet this criterion. Alternatively, suppose a study analyzes an employment training program. For it to meet ITS.5, researchers must determine (or provide instructions to determine) when those in the study begin program participation.

Note that ITS.5 should not be assessed if ITS.1 is not met, as a study must demonstrate that the timing of the intervention does not bias results before an analysis of who determines the timing of the intervention becomes relevant.

REFERENCES

Anderson, T. (1984). *An introduction to multivariate statistical analysis* (second edition). Hoboken: Wiley-Interscience.

Angrist, J., & Krueger, A. (1999). Empirical strategies in labor economics. In A. Ashenfelter & D. Card (Eds.), *Handbook of labor economics*, vol. 3. Amsterdam, Netherlands: Elsevier.

Ashenfelter, O. (1978). Estimating the effects of training programs on earnings. *Review of Economics and Statistics, 60*,47–57.

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics, 119*(1), 249–275.

Bound, J., Jaeger, D., & Baker, R. (1995). Problems with instrumental variables estimation when the correlation between instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, *90*(430), 443–450.

Cragg, J., & Donald, S. (2006). Inferring the rank of a matrix. *Journal of Econometrics*, *76,* 223–250.

Donald, S., & Lang, K. (2007). Inference with difference-in-difference and other panel data. *Review of Economics and Statistics, 89*(2), 221–233.

Greene, W. (2011). *Econometric analysis*, (seventh edition). New York: Pearson.

Hahn, J., & Hausman, J. (2002). A new specification test for the validity of instrumental variables. *Econometrica, 70*(1) 161–189.

Johnston, J., & DiNardo, J. (1997). *Econometric methods*, fourth edition. New York: McGraw-Hill.

Stock, J., & Yogo, M. (2005). Testing for weak instruments in linear IV regression. In J. Stock & D. Andrews (Eds.), *Identification and inference for econometric models: Essays in honor of Thomas J. Rothenberg* (pp. 80–108). Cambridge, England: Cambridge University Press.