Systematic Reviews in Education Research: When Do Effect Studies Provide Evidence?

Chris van Klaveren and Inge De Wolf
Systematic Reviews in Education Research: When Do Effect Studies Provide Evidence?¹

Chris Van Klaveren²

Inge De Wolf³

Systematic reviews often include studies based on rather ad hoc inclusion criteria that may lead to inconsistencies in the outcomes generated by these reviews. This study therefore derives a set of inclusion criteria from the theoretical literature on causal inference. It, furthermore, presents a list of study characteristics that reviews should minimally provide (if possible). By comparing the derived inclusion criteria in this study with the inclusion criteria used in several educational reviews, we conclude that the generated outcomes of systematic reviews frequently rely on rather incomplete and inconsistent inclusion criteria.

JEL Code: I20, C01.
Keywords: causal inference, systematic reviews, education.

¹We would like to thank Kristof De Witte, Hessel Oosterbeek, and Robert Slavin for their valuable remarks.
²The corresponding author is affiliated with Maastricht University, Top Institute for Evidence Based Education Research (TIER), P.O. Box 616, 6200 MD Maastricht, The Netherlands. Email: cp.vanklaveren@maastrichtuniversity.nl.
³The author is affiliated with the University of Amsterdam, Top Institute for Evidence Based Education Research (TIER), Email: i.dewolf@uva.nl.
1. Introduction

“Although purely descriptive research has an important role to play, often social science research is about cause and effect. A causal relationship is useful for making predictions about the consequences of changing circumstances or policies; it tells us what would happen in alternative (or counterfactual) worlds (Angrist and Pischke, 2009)”

Systematic literature reviews summarize the results of previous studies to inform the reader on the effectiveness of certain programs and give structure to the findings of larger amounts of empirical studies that focus on the same research topic. By putting the emphasis on causality, these reviews recognize that policy makers and scientist should be informed about causal relationships to make (more) accurate predictions about the consequences of changing circumstances or policies (Angrist and Pischke, 2009). For this purpose a distinction is being made between studies that focus on causal inference and other studies (i.e. correlational studies). Biomedical and psychological reviews frequently define consistent and clearly stated a priori inclusion criteria to make this distinction (see for example Slavin, 1984; 1986; 2008, and What Works Clearinghouse, 2008).

Currently, there is no predefined set of inclusion criteria for writing a systematic review on causal inference in economic and social sciences in general. This has, first of all, to do with the fact that many econometric evaluation approaches on causal inference were developed and applied in the last two decades (Imbens and Wooldridge, 2008). Examples are regression discontinuity, instrumental variable analysis, and matching approaches. These evaluation approaches are frequently used in economics, but also in other educational sciences (see, for example, Ou and Reynolds, 2010, and Greene and Winters, 2007). Each of these evaluation methods have their own pros and cons, as is illustrated by numerous methodological reviews: Rubin, 1974; Angrist and Kruger, 1999; Webbink, 2005; Blundell and Costa Dias, 2009; Imbens and Wooldridge, 2008; Angrist & Pischke, 2009; Imbens, 2009). Furthermore, there are few experiments in the economic and social science literature. Inclusion criteria in biomedical and psychological reviews are purely formulated with randomized controlled (field) experiments as a reference point, but these experiments are often not feasible in social sciences (see, for example, Cook, 2002, and Rodrik, 2008). The combination of the lack of randomized experiments and the larger variety in evaluation methods for causal effects, makes it more difficult to formulate consistent and transparent a priori inclusion criteria in economic reviews.

The inclusion criteria that are currently used in systematic economic reviews are not derived from the theoretical literature on causal inference. As a consequence, it is not clear how complete specific sets of inclusion criteria are, in the sense that it remains unclear if additional inclusion criteria should have been specified, or if the specified inclusion criteria are overly restrictive. In the former case, correlational studies may slip through the selection procedure imposing a bias on the review findings, and in the latter case, qualitative good and causal studies may not be considered in the review, which can also impose a bias on the research findings.
This study derives a set of inclusion criteria from the theoretical literature on causal inference and can be applied in systematic reviews to select and rate studies. By doing so, the underlying assumptions of the derived inclusion criteria are at all times clear. This study is complementary with currently existing biomedical and psychological studies that derive inclusion criteria from a fundamental discussion of theory (Slavin, 1984; 1986; 2008, and What Works Clearinghouse, 2008) because it (a) focuses on economic and other social sciences literature and (b) considers the econometric evaluation methods on causal mechanisms of the last two decades.

The second contribution is that it presents study characteristics that reviews should minimally provide to the reader (if possible) in order to be informative. These study characteristics are also derived from the theoretical literature on causal inference and can be used, together with the derived inclusion criteria, to select and rate the results of empirical studies at its true evidence.

Thirdly, this study examines if systematic review outcomes rely on complete and consistent inclusion criteria. For this purpose we compare the derived inclusion criteria in this study to the inclusion criteria used in 18 review studies that were published in the Review of Education Research in 2010.

Finally, this study is of practical value for policy makers, principals and teachers, as it provides them with an easy-to-apply list of rules that can be used to review empirical studies for evidence based policy or evidence based practise.

The theoretical framework used to derive the inclusion criteria from is inspired by Imbens (2009), who refers in his paper to conditions related to research methodology (first order conditions) and conditions related to analytical and data issues (second order conditions). We, additionally, define a group of third order conditions, which are related to representativeness, impact and usefulness of an estimated causal relationship. These third order conditions enfold external validity, the relevance of effect sizes of estimators, and the extent to which experimental outcomes are informative for implementing an intervention in practice. Based on these first, second and third order conditions, this study concludes that there are (only) four rather objective inclusion criteria that systematic literature reviews should apply.

This methodological focus of this study results in inclusion criteria that are methodological as well. There are, however, also other relevant inclusion criteria for systematic reviews that are related to the nature of the evaluated intervention. Slavin (2008), for example, formulated a list of conditions that interventions must satisfy in order to be informative. One of these conditions is that evaluation studies should only be included in reviews if the duration of the evaluated intervention is minimally 12 weeks. This condition gives more certainty that the reviewed interventions have substantial impact and are of practical use in education (Slavin, 2008).

We proceed as follows. Sections 2 to 4 discuss the first, second and third order conditions. From these conditions we derive (a) minimum quality standards that should be used as inclusion criteria and (b) study characteristics that reviews should show in order to be informative. The study characteristics indicate if a particular study is relevant in the ‘eye of the beholder’, but are not related to the quality of the study, in the sense that it is not related to the internal validity of the estimated impact. Section 5
compares the derived inclusion criteria to the applied inclusion criteria by 18 review studies published in the Review of Education Research. Finally, Section 7, concludes the main findings.

2. First order conditions

A missing data problem is underlying each evaluation study, which is nicely illustrated by the potential outcome model, introduced by Neyman (1990), Roy (1951); Rubin (1974, 1976) and Holland (1986). The model recognizes that there are two potential outcomes for each person\(^4\) to consider if we want to measure the effectiveness of an intervention. The intervention may represent a real intervention, such as a nation wide policy intervention or a summer school program, but it may also represent a characteristic, such as having a teacher credential. A person may either be subject to the intervention or may not, but no person can be in both states simultaneously. Since we never observe both potential outcomes at the same time it is not possible to measure the effect of the intervention directly. To the outcome not observed is generally referred to as the \textit{counterfactual outcome}.

Given that there are always counterfactual outcomes in evaluation studies, the effectiveness of an intervention can only be measured if there is a control group to which the outcomes of the persons subject to the intervention (the intervention group) can be compared to. This control group and the intervention group should have (on average) similar observable and unobservable characteristics that influence the outcome variable, such that outcome differences between the two groups are due to the intervention alone. The effect of the intervention on the outcome variable is then obtained by comparing the outcomes of the intervention group with those of the control group. It follows that the degree to which studies are able to measure how the outcome variable is causally related to the intervention (i.e. deliver evidence) depends on how well these studies are able to construct a proper control group.

This section discusses the methodological conditions that are related to how well studies are able to construct a proper control group. Because these methodological conditions are unrelated to data problems or analytical problems, and therefore always apply, we refer to them as first order conditions. We can distinguish three types of studies that construct the control group in a fundamentally different way:

\begin{enumerate}
  \item Randomized Controlled Trials.
  \item Observational Studies:
    \begin{enumerate}
      \item Quasi Experimental Studies,
      \item Other Observational Studies.
    \end{enumerate}
\end{enumerate}

\(^4\) Throughout the paper we conveniently refer to persons as the unit of analysis. Obviously the unit of analysis may also be countries, organizations, firms, classrooms, schools, teachers, employers etc.
Randomized Controlled Trials
Randomized controlled (field) trials (RCTs), assign persons (conditionally on certain characteristics) randomly to a control group and an intervention group. The randomization ensures that both groups are similar in both observable and unobservable characteristics, given that there are enough participants in the experiment. Differences in outcomes between the intervention and the control group can therefore be attributed solely to the intervention.

Fisher (1925) has shown, more formally, that randomization enables the researcher to quantify the uncertainty associated with the evidence for an effect of a treatment. Specifically, it allows for the calculation of exact p-values of sharp null hypotheses. These p-values are free of assumptions on distributions of outcomes, assumptions on the sampling process, or assumptions on the interaction between persons, solely relying on randomization and a sharp null hypothesis (Imbens, 2009).

Randomized (field) experiments are, however, frequently considered as not possible, uninformative or unethical. Randomization is for example rarely possible with specific macro-economic policies, such as trade, monetary, and fiscal policies (Rodrik, 2008). The results of randomized (field) experiments are, furthermore, often considered to be uninformative when experiments focus on a too specific population, when researchers aim at measuring a local average treatment effect, instead of an average treatment effect, and when experiments are conducted in a too controlled (or laboratory) setting. In these cases, generalization of the research findings might be problematic and the estimated effects may be too narrow to base policy on. It is also commonly remarked that randomized experiments are unethical because one random part of the research population is excluded from participation in a potentially effective intervention.

Quasi Experimental Studies
Quasi experimental studies (QES) are viewed as credible alternatives for studies that use data from RCTs, for example when RCTs are considered as not possible, uninformative or unethical. In QES, researchers mimic the randomized assignment of the experimental setting, but do so with non-experimental data (Blundell and Costa Dias, 2008). The intervention group includes self-selected persons (for example, students who join a summer school program) or persons who are selected by another process (for example, through a lottery design), along with a control group of persons. Since persons are not randomly distributed over the intervention and the control group, it follows that both groups may have different observable and unobservable characteristics. QES must therefore demonstrate that both groups are equivalent on observable characteristics and must assume that this equivalence is enough to ensure that both groups are also equivalent on unobservable characteristics. Obviously, the latter can not be verified, it can only be argued.

Imbens (2009) rightfully remarks that in a situation where one has control over the assignment mechanism, there is little to gain, and much to loose, by giving that up through allowing individuals to choose their own treatment regime. Randomization ensures exogeneity of key variables, where in a corresponding observational study one
would have to worry about their endogeneity. Hence, from a first order perspective randomized controlled trials are methodologically superior to observational studies.\(^5\)

But often there is no situation where the researcher has (full) control over the assignment mechanism. In situations where this happens, QES may not only serve as a credible alternative for studies that use data from RCTs, it may even be the best evaluation method at hand.

There are (roughly) five types of QES: (i) natural experiments, (ii) discontinuity design methods, (iii) matching methods, (iv) instrumental variable methods, and (v) difference-in-differences or similar identification methods that rely on panel data. Natural experiments exploit randomization to the intervention created through some naturally occurring event external to the researcher, i.e. forces of nature or a policy change (see, for example, Angrist and Evans, 1998). The line of argumentation of these studies is that persons are randomly allocated to an intervention and a control group due to an exogenous event, such that both groups are equivalent in both observables and unobservables.

Discontinuity design methods exploit `natural' discontinuities in the rules used to assign individuals to a treatment (see, for example, Lee, 2008). It uses the fact that in highly rule-based worlds, some rules are arbitrary and therefore provide good experiments (Angrist and Pischke, 2009). An example taken from Angrist and Pischke illustrates the discontinuity design method more intuitively. American high school students are awarded National Merit Scholarship Awards on the basis of PSAT scores, a test taken by most college-bound high school juniors, especially those who will later take the SAT. To answer the question if students who win these awards are more likely to finish college one could relate college completion rates to winning the award, but the problem is that students with higher PSAT scores are more likely to finish college. A discontinuity design overcomes this problem by comparing college completion rates of students with PSAT scores around the award cutoff. The discontinuity design therefore compare student groups in the neighborhood of the award cutoff to ensure equivalence in observables and assume equivalence on unobservables.

Matching methods do not use exogenous variation from natural events or natural discontinuities, but instead simulate a control group from the group of non-treated, and re-estabishes in this way the experimental conditions in a non-experimental setting. Matching methods rely on observable variables to account for selection issues, and it assumes that the selection on unobservables, that is not accounted for by observables, balances out between the intervention group and the control group (see, for example, Rubin, 1976). Matching methods require rich data, especially on a pretest or premeasurement. Otherwise, more complex matching methods are needed in other to evaluate causal mechanisms. De Witte and Van Klaveren (2012), for example, compare drop out rates in two Dutch cities (Amsterdam

\(^5\)In the presence of a Hawthorne-effect it may, however, be that RCTs are not superior to QES. The Hawthorne effect is a form of reactivity whereby subjects improve or modify an aspect of their behavior because they are being studied, and not because of the particular incentive they receive during the experiment (see Franke and Kaul, 1978). While in RCTs subjects often are aware of the fact that they are being studied, this is not always the case in QES.
and Rotterdam) and argue that both student populations are very different in characteristics that may influence the observed drop out rates. To address this problem they perform an interactive matching approach and match a sample of Rotterdam students to the best look-alike sample of students in Amsterdam and repeat this procedure a thousand times. In this way they reconstruct the drop out rate distributions for the student populations in both cities given that the underlying population characteristics are comparable. The matching design therefore enforces equivalence on those observables included in the matching procedure and it is assumed that this is enough to enforce equivalence on unobservables.

Instrumental variable methods (IV) differ from the evaluation methods above, because they focus on the selection on the unobservables. IV requires that there is at least one instrument exclusive to the participation rule. The identifying assumption in IV is that the instrument is only related to the assignment mechanism, and not directly to the outcome variable of interest (see Heckman, 1997). Hence, the instrumental variable is assumed to affect the participation rule, which in turn affects the outcome variable. A well known IV study is that of Angrist (1990) who studies the effect of serving in the Vietnam war on the earnings of men. Angrist uses the fact that men were selected based on a lottery draft. This lottery draft was randomly determined and men with a lower draft lottery were more likely to participate in the military. The IV method predicts, in this case, the probability of participation in the military (controlling for observables that affect earnings) and then relates this predicted value to the earnings of men (controlling for observables that affect earnings). Hence, the lottery draft is assumed to affect the participation rule, which in turn affects the earnings of men.

The problem with measuring the effect in the Vietnam War study, is that participation in the military is not independent from unobserved factors that affect earnings (even if the researcher controls for observables). In one specific case, the identifying IV assumption is satisfied by construction: randomized (field) trials, where persons are not treated consistent with how they were assigned to the treatment- or control group (i.e. there is non-compliance). In this specific situation, the observed participation can be instrumented by how persons were assigned to the treatment. Because the assignment of persons to the treatment was random, it is true by construction that the instrument is only related to the assignment mechanism, and not directly to the outcome variable of interest.

The fifth type of quasi-experimental study is referred to as difference-in-differences (DiD) and rely on panel data. These studies make use of data with a time or cohort dimension to control for unobserved but fixed omitted variables (Angrist and Pischke, 2009). DiD approaches compare persons who are subject to an intervention during a certain time period to other persons who are not subject to the intervention during the same time period. DiD studies must, first of all, show that the intervention and the control group have similar observed characteristics, as well as similar outcomes of interest, before the intervention took place. Secondly, these studies can only argue that the intervention was effective if the measured outcome changes at the moment the intervention group becomes subject to the intervention, while at the same time no such change is observed for the control group.
DiD studies generally use a pre- and post-measurement for both the control and the intervention group, but unlike RCTs, persons are not randomly assigned to both groups. To illustrate the intuition and the shortcomings of DiD we shortly discuss a study of Card and Krueger (1994) who performed a DiD analysis to examine how employment is affected by a dramatic change in the New Jersey state minimum wage, on April 1, 1992. Because fast food restaurants employ many people at the minimum wage, Card and Krueger (1994) collected employment data of fast food restaurants in New Jersey and Pennsylvania, in February 1992 and again in November 1992. The minimum wage in New Jersey increased from $4.25 to $5.05, while the minimum wage in Pennsylvania was $4.25 throughout this period. When they compared the change in employment in New Jersey to the change in employment in Pennsylvania, and controlling for relevant (state) characteristics, they found that the increase in wage increased employment, instead of reduced employment.6

The effect found by Card and Krueger (1994) is rather unexpected and it places some doubts on the (internal) validity of the estimated effects. The main problem with DiD analysis is that it is usually very difficult to control for unobserved factors, while (especially for studies more on the macro level) these unobserved factors will likely affect the empirical outcomes. Frequently, DiD studies do not show the trend of the outcome variable, separately for the intervention and the control group, for a long period before and after the change or intervention took place.

Given the frequent lack of longitudinal data, diff-in-diff studies usually do not examine (1) how outcomes fluctuated over time and (2) how these fluctuations influence the estimated effect. In addition the (counterfactual) trend behavior of the intervention and the control group should be the same (Angrist and Pischke, 2009), which is often difficult to argue. DiD analysis is, however, intuitive and the problems mentioned above either also occur with other QES (i.e. equivalence on unobservable cannot be shown) or comes from the fact that longitudinal data is used, which is not a drawback by itself.

A clear distinction should, however, be made between DiD models and other panel data approaches. Panel data studies generally do not define a clear control group and, as a consequence, it is not possible to compare observable characteristics between the intervention and the control group. Moreover, these studies tend to examine how many characteristics influence one single outcome variable. With respect to the internal validity of estimators this is problematic, because the selection mechanism of having a certain characteristic may vary per characteristic. Therefore panel studies that do not adopt a DiD strategy should not be classified as QES and should not be included in a systematic literature review as the empirical results generated by these studies cannot be interpreted causally.

QES differ from RCTs because persons are not randomly assigned to the treatment. Nevertheless, these studies may produce unbiased estimates of the effect of an intervention, and therefore deliver evidence. The quality of the evidence delivered by QES depends on wether the underlying model assumptions are valid, and the main

6 The example is taken from Angrist and Pischke (2009)
drawback is that equivalence on unobservables between the control and intervention group cannot be shown. It goes beyond the scope of this article to discuss each of the quasi experimental evaluation methods in more detail. Thorough surveys of quasi experimental designs are given, among others, by Webbink (2005), Blundell and Costa Dias (2009), Imbens and Wooldridge (2008), Angrist and Pischke (2009) and Imbens (2009).

**Correlational Studies**
Quantitative correlational studies do not construct a (valid) control group or do not focus on the selection into a certain treatment. This makes these studies fundamentally different from QES and RCTs. Correlational studies examine how observed variation in the measured outcome variable is explained by observed variation of several characteristics and one of these characteristics indicates if a person is treated or not. The focus of these studies is therefore not on the internal validity of estimators and observed outcome differences between the intervention and the (undefined) comparison group may have been caused by the intervention, but also by other (un)observed factors. It follows that correlational studies may produce biased estimates of the effect of the intervention and should not be included in systematic reviews.

It happens far too often that the empirical findings of correlational studies are interpreted as if they represent a causal effect, even though they merely represent a correlation. Moreover, and even worse, many policies are implemented on the basis of correlational studies, while these studies are not informative about the impact of effect of policy measures.

The internal validity of estimators does not play an important role in correlational studies. As a consequence, it is unlikely that correlational studies provide internally valid estimators and that the findings of these studies can be causally interpreted. *The first inclusion criteria is therefore that correlational studies should not be included in systematic literature reviews.*

Based on the discussion of the first order conditions, we conclude that randomized controlled (field) experiments are at the top of the hierarchical ladder when it comes to measuring causal effects. For this reason, randomized control trials are often viewed as the ‘golden evaluation standard’ (see also the ‘Maryland Scale of Scientific Methods, MSSM), but the term golden is misleading as it suggest that findings from randomized controlled trials should always be preferred to the findings from other studies, including those of quasi-experimental studies. In Section 3, where we discuss the second order conditions, we point out that this view is incorrect. QES may provide estimates as reliable as the estimates of the RCT, as long as the underlying model assumptions are valid.

Experimental data are, moreover, often not available and researchers usually do not have control over the assignment mechanism. In these situations QES may not even be a credible alternative for studies that use data from RCTs, it may simply be the best evaluation method at hand. At the same time, and purely from a methodological perspective, it holds that RCTs are superior to QES, because QES must *demonstrate* that the intervention and control group are equivalent on observable characteristics and
must assume that this equivalence ensures equivalence on unobservable characteristics. The latter assumption cannot be formally tested. Also RCTs should demonstrate equivalence of observables between the control group and the intervention group to show that the randomization was successful. The second inclusion criteria therefore is that RCTs and QES must show the equivalence on observables between control and intervention group. This condition includes that the IV studies should make plausible that the instrument is only related to the assignment mechanism, and not directly to the outcome variable of interest.

Differences in the methodological approach adopted may lead to differences in the estimated effect. It is, for example, difficult to check if a study uses a valid instrument, and the decision on whether a study uses a valid instrument is at least subjective. It is therefore important that systematic reviews verify if the estimated effects depend on the chosen methodological approach. Systematic reviews should therefore present information on the used methodological approach of the reviewed study.

3. Second order conditions

Second order conditions are conditions that are related to analytical and data issues of the reviewed studies. Data issues or analytical deficiencies may be independent of the research methodology used, but can bias the estimated effectiveness of an intervention.

In Table 1 we present a list of second order problems that may occur, like, attrition, missing values, (non-)clustered standard errors, learning effects, equivalence of the intervention and control group, problems with the validity of the used instrument to identify the effect of the intervention and problems with the power of the analysis (i.e. small sample bias). Some of these second order deficiencies are relatively easy to address. For example, in randomized experiments it may be difficult to reject the null hypothesis that the intervention is not effective due to a small sample size. By reporting the power of the test, authors can show that the estimated effect is large enough to reject the null hypothesis, even though the sample size is small.

Unfortunately, not all second order deficiencies can be easily addressed. For example, the validity of an instrument cannot be formally shown, it can only be argued by the authors. Sometimes it is obvious that an instrument is valid (e.g. participation status is instrumented by the random assignment status when there is little non-compliance), but it may not be so obvious in other cases (e.g. the Vietnam lottery example of Section 2). Subjective discussions on the validity of instruments are informative, but are not useful for defining inclusion criteria that determine when studies are to be included in a literature review.
<table>
<thead>
<tr>
<th>Data Problems</th>
<th>Analytical Problems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small sample size</td>
<td>Standard errors (not) clustered</td>
</tr>
<tr>
<td>Attrition</td>
<td>Same Pre- and Post-test (learning effects)</td>
</tr>
<tr>
<td>Missing values/information</td>
<td>Equivalence of intervention and control group not shown</td>
</tr>
<tr>
<td>No common support</td>
<td>Nonvalid Instrument</td>
</tr>
<tr>
<td>Contamination</td>
<td>Power of the test not calculated</td>
</tr>
<tr>
<td>Cross-over effects</td>
<td></td>
</tr>
</tbody>
</table>

The difficulty with second order deficiencies is that it is generally not clear how they bias the estimated effectiveness of an intervention. It is possible that a well designed QES provides a more accurate estimate than a methodologically superior RCT, due to more (unobserved) second order problems in the RCT-study. The difference between the estimated and true effectiveness can only be determined if the true effect of the intervention is known, and only then it would be possible to examine which studies provided the most accurate estimates. It follows that second order deficiencies and their unknown consequences for the estimates make it impossible to rank studies purely based on the chosen evaluation method. To judge the quality of evidence that is delivered by the included studies, it is necessary to evaluate and report data and analytical problems in systematic literature reviews.

To minimize the effect of second order problems we propose a third and a fourth inclusion criteria. The third inclusion criteria is that studies must be accepted for publication in international and peer-reviewed journals or must be published as a chapter in a peer-reviewed thesis. The intuition underlying this inclusion criteria is that reviewers, editors, promtors and committee members recognize analytical and data problems such that these problems are better adressed.

This criteria is, however, controversial due to the well known publication bias that may bias the estimated effectiveness of an intervention. It has been shown that studies are more likely to be published if the empirical findings suggest that an outcome variable is positively and significantly influenced by the intervention studied (Slavin, 2008). Considering only published and peer-reviewed studies in systematic literature reviews would then impose an upward bias on the estimated effectiveness of the intervention studied. The third inclusion criteria therefore implies that the problem of having analytical and data deficiencies is substituted for the problem of having publication bias. A major advantage of this inclusion criteria, however, is that the possible impact of publication bias is verifiable, while the possible impact of second order deficiencies are not verifiable. To have an idea of the impact of publication bias it is important that systematic literature reviews verify how findings of non-published studies that satisfy the first two inclusion criteria compare to the findings of the studies that are included in the systematic literature review. Differences between the findings of these unpublished studies and the studies included in the systematic literature review are then either caused by publication bias or by second order deficiencies. The larger this difference is, the more sceptical one should be about the reliability of the estimated effectiveness of the intervention.
It is noteworthy that small sample experiments in medicine frequently find huge effect sizes and it is often reasoned that publication bias is a more serious problem in small sample research than in studies with large samples (Givens and Tweedie, 1997, Sterne, Gavaghan and Egger, 2000, Rothstein, Sutton and Borenstein, 2005). Therefore, the sample size of each study should be reported in a systematic literature review.

The fourth inclusion criteria is that the focus of the literature review must be congruent to the focus of the studies included in the review. Studies often control for several characteristics that may affect the outcome variable, besides the intervention. The first and second order problems are usually not addressed for these control variables, but are only addressed for the evaluated intervention. Studies that are interested in the effectiveness of the intervention, frequently also show and discuss how different control variables affect the outcome variable. The estimates of these control variables can obviously not be interpreted as evidence.

4. Third order conditions

Third order conditions are related to the representativeness or usefulness of the empirical findings of studies that are included in systematic literature reviews (i.e. the findings of studies that satisfy the first four inclusion criteria). These conditions enfold the issues of external validity, effect size and informativeness for the actual implementation of a program. We subsequently discuss each of these three issues, starting with external validity.

RCTs and QES value the internal validity of estimates more than the external validity of estimates, and reliable estimates for specific subpopulations are therefore preferred to representative but unreliable estimates for the overall population. It is, however, unclear how representative findings for one subpopulation are for another (sub)population. To give an extreme example, if the findings of a perfectly conducted randomized controlled trial in KwaZulu-Natal (province of South Africa) indicate that a summer school program was effective, then this (obviously) does not mean that a similar summer school program will be effective in New York as well. The value of empirical evidence thus depends on what the evidence is needed for and how it will be used (see also Rodrik, 2008, who has a similar argument).

Even though there is a debate on whether internal validity should be preferred over external validity (see Banerjee and Duflo, 2008), this discussion is not so relevant for the inclusion of studies in systematic literature reviews. Internally valid estimators are estimators that are reliable given the specific conditions under which the study was performed. It applied that the representativeness of estimates for other populations is in the 'eye of the beholder'. A summer school evaluation for the Netherlands is likely to be more representative for New York than a summer school evaluation for KwaZulu-Natal, but policy makers and researchers in New York can probably judge best on how representative the empirical findings are. It therefore implies that systematic reviews should present information on the research population used in the included studies, but no inclusion criteria can be formulated on the basis of external validity.

The second issue concerns the relevance of effect sizes in systematic literature
reviews. Literature reviews (and meta-analysis) often conveniently use standardized effect sizes to label the empirical evidence as weak, moderate or strong.\(^7\) These effect sizes quantify the impact that an intervention has on the outcome variable, and enables comparison across studies and across interventions. The latter gives information on which intervention is the more effective one.

Reviews usually tend to focus on 'what works' and associate larger effect sizes with stronger evidence. A consequence is that studies with negative or zero effect sizes are less often mentioned in review reports. The 'what works'-paradigm implies a risk for falsification (ignoring Popper’s ‘black swan’) and may give a misleading message about the effectiveness of a certain intervention to policy makers and educational practitioners. Information on the ineffectiveness of interventions is equally informative than information on the effectiveness of interventions and systematic reviews should therefore include studies irrespective of the effect size they find.

Two practical problems with effect sizes are that (a) studies frequently do not report them (nor the statistics needed to calculate one) and (b) that there is no consensus about what an effect size represents and how it should be calculated. To illustrate the latter: effects sizes are defined using Cohen's \(d\) (Cohen and Cohen, 1983), or Hedges' \(g\) (Hedges, 1981), or are defined as the proportion of a standard deviation by which the experimental group outperformed the control group (adjusting for any pretest differences) (www.bestevidence.org). An advantage of reporting effect sizes is that empirical findings of different empirical studies can be compared, but this advantage disappears when different effect-size definitions are used. Systematic reviews should thus report effect sizes and the definition of these effect sizes.

A third issue relates to the similarity between the evaluated programs and the programs that people plan to implement (on a large scale). Review studies should therefore provide information about cost effectiveness and upscaling possibilities. The importance of this issue can be illustrated by the following example. Suppose that a summer school program was evaluated by a well designed RCT. To guarantee that the experiment was properly conducted each student in the summer school program was guided by a university student. The experimental results show that the summer school is effective, however the effectiveness may be caused by the intensive personal guidance rather than by the summer school program. If the summer school program would be implemented on a large scale, such intensive guidance would be practically undoable and too costly. Hence the experiment practically has no value because the evaluated program will never be implemented in practice.

Again, judgements on cost-effectiveness, implementation and upscaling is in the eye of the beholder (i.e. different persons may have different opinions on what, for example, cost-effectiveness is). Systematic reviews should therefore be as informative as possible on these issues, but the inclusion criteria of systematic reviews should be independent from these issues because the internal validity of the estimated effectiveness is independent from these issues.

---

\(^7\) See, for example, the ‘Best Evidence Encyclopedia’ (www.bestevidence.org).
5. Applying the Inclusion Criteria in Practice

This section examines to what extent systematic literature reviews already apply the four defined inclusion criteria, and if systematic reviews provide information about external validity, effect sizes and cost-effectiveness. For this purpose we examine the applied inclusion criteria and the information provided by a series of articles in Review of Educational Research. This journal specialized in systematic literature reviews and publishes articles that come from various social science disciplines, namely pedagogy, sociology, psychology and economics. In particular we examined all review studies that were published in 2010 in this journal.

By comparing the derived inclusion criteria in this study with the inclusion criteria used by the reviews published in Review of Educational Research we obtain more insight on (1) whether there is consensus about the inclusion criteria that should be used, and on (2) the completeness and consistency of the applied inclusion criteria in systematic reviews.

Under the assumption that the derived inclusion criteria in this study are correct, it furthermore gives an impression of the extent to which systematic reviews focus on ‘evidence’ (i.e. internal validity) and if these reviews provide information that is relevant for the implementation of (policy) interventions.

Reviews in Review of Educational Research

18 review studies were published in Review of Education Research in 2010. The inclusion criteria used in these studies varied from 1 to 7, and studies used 3.7 inclusion criteria on average. Appendix A provides a detailed description of the applied inclusion criteria by these 18 reviews.

Table 2 shows if the published review studies applied the inclusion criteria derived in this study. The columns indicate each inclusion criteria that systematic literature reviews should adopt:

I. Include only RCTs and QES.
II. Include only RCTs and QES that show the equivalence on observables between of the control and the intervention group.
III. Include only published and peer-reviewed studies (or theses)
IV. Include only studies that focus specifically on the topic reviewed.

The table indicates Y(es) if a study satisfies the inclusion criteria, and N(o) otherwise. The bottom row of Table 2 indicates the percentage of studies that satisfied the first, second, third and fourth inclusion criteria, respectively.

The table clearly shows that none of the reviews apply all four inclusion criteria, which indicates that it is not common practice to apply the four derived inclusion criteria. Two thirds of the reviews included only studies that specifically focus on the review topic (i.e. inclusion criteria IV.). These reviews generally do not provide an
argument for applying this inclusion criteria; it is just applied without any argument, or it is viewed as a logic consequence of the search strategy that authors used in their review. Those review studies that gave an argument for not applying the fourth inclusion criteria mentioned that this broadened the number of studies in the review. Thereby, these studies attach more value to increasing the number of studies than to including studies that better address data and analytical problems.


<table>
<thead>
<tr>
<th>Authors</th>
<th>Pages</th>
<th>I.</th>
<th>II.</th>
<th>III.</th>
<th>IV.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bowman</td>
<td>4-33</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Cooper et al.</td>
<td>34-70</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Achinstein et al.</td>
<td>71-107</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Superfine</td>
<td>108-137</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Bowers et al.</td>
<td>144-197</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Nguo Ng et al.</td>
<td>180-206</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Adesope et al.</td>
<td>207-245</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Mills</td>
<td>246-271</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Crede et al.</td>
<td>272-295</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Marulis et al.</td>
<td>300-335</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Cavagnetto</td>
<td>336-371</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Rakes et al.</td>
<td>372-400</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Patall et al.</td>
<td>401-437</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Goldrick-Rap</td>
<td>437-469</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Holme</td>
<td>476-526</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Johnson</td>
<td>527-575</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Bohnert et al.</td>
<td>576-610</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Lovett</td>
<td>611-638</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>17%</td>
<td>17%</td>
<td>28%</td>
<td>67%</td>
</tr>
</tbody>
</table>

Five of the 18 review studies (28 percent) include only peer reviewed articles. These reviews all argue that quality of peer reviewed articles is guaranteed, while there are no guarantees in the ‘gray’ literature. The main arguments of review studies to include also non-published studies are (a) to avoid possible selection bias and (b) because this inclusion criterion reduces the number of studies that can be selected. Section 3 explains that analytical and data deficiencies are reduced by including only published and peer-reviewed studies, but that this choice, at the same time, increases the likelihood of having publication bias. Review studies should therefore compare the findings of unpublished studies and published studies (given that these studies satisfy the other formulated inclusion criteria). Differences in research findings are then either caused by publication bias or by second order deficiencies, and the larger this difference the more sceptical one should be about the reliability of the estimated effectiveness of
the intervention. One could, for example, define a reliability index that divides the estimated impact of published studies by the estimated impact by non-published studies. This reliability index would then indicate 1 if the estimated impact by published studies is similar to that of non-published studies.

Finally, the table shows that only 3 review studies (17 percent) exclude correlational studies. Review studies that exclude correlational studies also include only (quasi)experimental designs with well matched treatment and control groups. It seems to be the case that if studies satisfy inclusion criteria II, they automatically satisfy inclusion criteria I. Review studies that do not satisfy the first two inclusion criteria do not provide evidence because the reviewed studies do not focus on the internal validity of estimators.

Table 2 suggests that 3 of the 18 review studies potentially provide evidence because they satisfy the first two inclusion criteria. None of these studies include only published studies which reduces possible publication bias, but which, at the same time, increases the likelihood that data and analytical problems are not properly addressed. Two of the three more causally oriented literature reviews include studies that do not specifically focus on the topic reviewed. This may bias the review findings because data and analytical problems may not be properly addressed (see section 3).

Table 3 shows whether the 18 review studies provide information on external validity, effect sizes and cost-effectiveness or up-scaling possibilities. Section 4 argues that this information is essential for making an informative choice between different potentially effective interventions and the implementation of these specific interventions.

The table shows that it is not common practice for review studies to provide this information. Less than 50 percent of the review studies report effect sizes. Some reviews mention that they only report effect sizes if possible, which points to the fact that effect sizes are not always given or cannot always be calculated from the study reviewed. The absence of effect sizes makes it more difficult to compare differences in the measured intervention impact of the reviewed studies.

Only 7 of the 19 review articles (39%) described the research populations of the reviewed studies. As a consequence, the value of the delivered empirical evidence by these review studies remains unclear, because, in the end, this value depends on what the evidence is needed for and how it will be used.

Finally, Table 3 shows that information on the cost-effectiveness and up-scaling possibilities is never presented in the reviews (last column). One of the probable reasons is that the reviewed studies do not provide this information, as is suggested by Marulis et al. (2010).
Table 3. Information provided in literature reviews in Review of Education Research 2010

<table>
<thead>
<tr>
<th>Authors</th>
<th>pages</th>
<th>external validity</th>
<th>effect sizes</th>
<th>cost-effectiveness/implementation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bowman</td>
<td>4-33</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Cooper et al.</td>
<td>34-70</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Achinstein et al.</td>
<td>71-107</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Superfine</td>
<td>108-137</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Bowers et al.</td>
<td>144-197</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Ngan Ng et al.</td>
<td>180-206</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Adesope et al.</td>
<td>207-245</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Mills</td>
<td>246-271</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Crede et al.</td>
<td>272-295</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Marulis et al.</td>
<td>300-335</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Cavagnetto</td>
<td>336-371</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Rakes et al.</td>
<td>372-400</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Patall et al.</td>
<td>401-437</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Goldrick-Rap</td>
<td>437-469</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Holme</td>
<td>476-526</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Johnson</td>
<td>527-575</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Bohnert et al.</td>
<td>576-610</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Lovett</td>
<td>611-638</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
</tbody>
</table>

Total: 39% 44% 6%

6. Concluding remarks

Systematic reviews that focus on internal validity generally include studies based on rather ad hoc inclusion criteria which may lead to inconsistencies in the outcomes generated by these reviews. Currently there is no predefined set of inclusion criteria and there are two reasons for this. First of all, many econometric evaluation approaches on causal inference were developed in the last two decades and, second of all, there are often few experimental studies. The formulation of consistent and transparent inclusion criteria is therefore difficult.

The objective of this study is to derive a set of inclusion criteria from the theoretical literature on causal inference. Thereby this study contributes to the literature by providing more consistent and transparent inclusion criteria which should increase the extent to which systematic reviews provide evidence. The second contribution of this study is that it presents study characteristics that reviews should minimally provide to the reader (if possible) in order to be informative that also follow from the methodological discussion on causal inference. The presented inclusion criteria and study characteristics are, furthermore, of practical value for policy makers, principals and teachers, as it provides them with an easy-to-apply list of rules that can be used to review empirical studies for evidence based policy or evidence based practice.
This study derives the following set of inclusion criteria from the theoretical literature on causal inference:

I. Include experimental or quasi-experimental studies only.

II. Include experimental or quasi-experimental studies that show the equivalence on observables of the control and the intervention group.

III. Include only published and peer-reviewed studies (or theses).

IV. Include only studies that focus specifically on the topic reviewed.

The characteristics that systematic reviews should minimally provide (if available) in order to be informative are:

- **[Identification]:** The adopted identification method
- **[external validity]:** The research population, conditions and publication year;
- **[effect size]:** The (standardized) effect size(s) and type of effect size;
- **[Implementability]:** Information on cost-effectiveness and upscaling possibilities.
- **[publication bias]:** A comparison of the findings of published and non-published studies that satisfy inclusion criteria I, II and IV.

These characteristics are not related to the internal validity of the estimated effect and are, therefore, not considered as inclusion criteria. However, showing these characteristics is extremely important for a better understanding of causal mechanisms and for using the review results for evidence based policy.

Finally this study examined to what extent systematic literature reviews already apply the four defined inclusion criteria and show the study characteristics indicated by this study. For this purpose we reviewed all 18 published review studies in Review of Education Research in 2010.

The inclusion criteria seem minimal and straightforward, yet only three of the 18 review studies satisfy the first two inclusion criteria. None of these three review studies included only published studies and 33 percent of the review studies included studies that do not focus specifically on the topic reviewed. The comparison indicates that it is not common practice to apply the four inclusion criteria derived in this study. Consequently, the results generated by these studies provide no causal evidence, but provide correlations which cannot be used by policy makers for evidence based policy.

With respect to the reported study characteristics we conclude that less than 50% of the review studies report information on effect sizes or external validity and only 1 study reported information concerning the implementability. The latter result is likely because the studies selected in review studies usually do not report information on implementability. As a consequence, the policy value of the generated result often remains unclear, because, in the end, this value depends on what the evidence is needed for and how it will be used.

Finally, we would like to discuss two issues with respect to the derived inclusion
criteria in this study. First of all, the choice to include only published and peer-reviewed studies in systematic reviews is controversial. It may impose a publication bias and at the same time it reduces the probability that the empirical results are driven by analytical and data deficiencies. Therefore, we added the study characteristic that results of published and unpublished studies that satisfy inclusion criteria I, II and IV should be compared. One should be more sceptical about the reliability of the review findings as the difference between the outcomes of published and unpublished studies becomes larger.

A second issue is that there are frequently no or few experimental or quasi-experimental studies available. Applying the inclusion criteria derived in this study would mean that systematic literature reviews can select no or only a few. At this point it is relevant to recognize the main objective of this study: deriving a set of consistent and transparent inclusion criteria from the theoretical literature on causal inference. If the best evidence is delivered by non-causal studies, then our study does not indicate that reviewing these studies is uninformative. Instead it suggests that systematic reviews that evaluate the outcomes of these studies are uninformative with respect to the underlying causal mechanism. In our opinion, these literature reviews should mention explicitly that the review outcomes may not be suitable for evidence-based policy making.

References


### Appendix A. Inclusion criteria used in review articles in the Review of Education, edition 2010

<table>
<thead>
<tr>
<th>Paper by</th>
<th>Pages</th>
<th>Inclusion criteria</th>
</tr>
</thead>
</table>
| Bowman             | 4-33  | - participants were undergraduate students or were reporting about their previous undergraduate experiences in the United States,  
|                    |       | - at least one independent variable measured a college diversity experience,  
|                    |       | - the DV measured cognitive skills or tendencies, and  
|                    |       | - statistics regarding the magnitude of the effect were provided.                                                                                                                                                                                                                         |
| Cooper et al.      | 34-70 | - the study had to have focused on the difference between kindergarten programs that operated on a half-day schedule versus a full-day schedule on a measure of student academic achievement or readiness, some other measure of student development or well-being, or some measure of classroom process.  
|                    |       | - no studies that compared HDK to FDK that met on alternating school days.  
|                    |       | - studies had to study kindergarten programs based in the United States or Canada.  
|                    |       | - no studies that intentionally confounded the FDK variable with another instructional intervention.  
|                    |       | - outcomes measured at the end of the kindergarten year or the beginning of first grade.  
|                    |       | - enough information to permit the calculation of an estimate of the relationship between the length of the kindergarten day and the outcome measure.                                                                                                                                                                      |
| Achinstein et al.  | 71-107| - reported an empirical study,  
|                    |       | - adhered to the American Educational Research Association’s (2006) published standards for reporting on empirical social science research, and  
|                    |       | - informed an important aspect of retention of new teachers of color, such as retention rates and turnover factors.                                                                                                                                                                                                 |
| Superfine          | 108-137| - directly applied to the focus of this review  
|                    |       | - authoritative nature of the literature (i.e., publication in a peer-reviewed journal, publication at a major academic press, or including extensive references to such sources in the case of law review articles), and  
|                    |       | - its relevance to issues of centralization and decentralization of authority.                                                                                                                                                                                                                   |
| Bowers et al.      | 144   | - Published in English, reporting on research carried out in an alphabetic orthography  
|                    |       | - Investigated instruction with elementary school students,  
|                    |       | - Investigated instruction about any element of oral or written morphology,  
|                    |       | - At least one third of the instruction was focused on morphology, based on the intervention description,  
|                    |       | - Reported literacy outcome measures with means and standard deviations for comparison,  
|                    |       | - Used either an experimental and control or comparison group or a training group with pre- and posttests using measures that could be compared to established norms  
| Ngan Ng et al.     | 180-206| - no conference papers, book reviews and unpublished materials.  
|                    |       | - no chapters that reported on sets of data in published studies that we had previously reviewed.                                                                                                                                                                                                 |
| Adesope et al.     | 207-245| - Bilingual participants were reported to be equally (or almost equally) proficient in two languages. Bilinguals with learning disabilities or other cognitive disabilities were excluded.                                                                                                                                                                                                 |
- They had an experimental group of bilingual participants and a control group of monolingual participants.
- Measured outcomes (cognitive benefits) were clearly reported.
- They reported sufficient data to allow for effect size calculations.
- They were publicly available, either online or in library archives.

| Mills      | 246-271 | timeframe of 1999 through 2009
|            |         | observational studies
| Crede et al.| 272-295 | report correlations between either (a) class attendance and college GPA or (b) attendance in a particular class and the grade obtained in that class or allowed computation of either of these relationships.
|            |         | no laboratory studies or attendance in high school or primary school classes.
|            |         | not if only statistically significant findings were reported
|            |         | no articles where the grade in a class was based, in part, on class attendance.

| Marulis et al.| 300-335 | the study included a training, intervention, or specific teaching technique to increase word learning;
|              |         | a (quasi)experimental design was applied, incorporating one or more of the following: a randomized controlled trial, a pretest–intervention–posttest with a control group, or a postintervention comparison between preexisting groups;
|              |         | participants had no mental, physical, or sensory handicaps and were within ages birth through 9;
|              |         | the study was conducted with English words, excluding foreign language or nonsense words (to be able to make comparable comparisons across studies); and
|              |         | outcome variables included a dependent variable that measured word learning, identified as either expressive or receptive vocabulary development or both.

| Cavagnetto | 336-371 | only articles that established a connection between argument and scientific literacy or the nature of science were used.
|            |         | articles reporting on research associated with argument-based interventions or a component of the intervention (articles from peer-reviewed practitioner journals were not included),
|            |         | articles that offered a clear description of the intervention, and
|            |         | articles reporting on interventions designed for K–12 students.

| Rakes et al. | 372-400 | the intervention had to target the learning of algebraic concepts, regardless of the title of the classes being examined.
|             |         | the intervention had to involve a method for improving learning as measured by student achievement
|             |         | employ an (quasi)experimental design with a comparison group.
|             |         | the comparison group had to have received the “usual instruction.

| Patall et al. | 401-437 | focus on the differences between students attending ED or EY schools and students at traditional day or year schools,
|             |         | a naturally occurring measure of number of school days or hours in the school day and its relationship to student achievement, or
|             |         | a description of programs that implemented changes in the length of the year or length of the day.
|             |         | appeared in 1985 or after.
|             |         | programs based in the United States or Canada.
|             |         | no studies examining the effect of full-day versus half-day kindergarten

| Goldrick-Rap | 437-469 | they used quantitative or qualitative methods that could rigorously address the research questions, and
|            |         | quantitative studies needed to produce findings that could reasonably be generalized beyond the sample to the larger population of community college
students.
- sufficient and appropriate data, following a research design that made it possible to answer the questions posed and a transparent and defensible approach to sampling and data collection.

<table>
<thead>
<tr>
<th>Author</th>
<th>Pages</th>
<th>Selection Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holme</td>
<td>476-526</td>
<td>relevance: only studies that pertained to the issue of exit testing for a high school diploma and that were relevant to our four central questions. We excluded studies that focused on students with disabilities and studies that focused on measurement and validity issues related to exit exams.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>empirical nature: both qualitative and quantitative studies that reported original research rather than the research of others, prior research, or opinion.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>scholarship quality, studies must have (a) provided a clear and well-supported logic of inquiry, (b) adequately described the design and sources of evidence, (c) used sources of evidence that were appropriate to the scope of the questions, (d) described the analysis procedures, and (e) described the warrants for conclusions.</td>
</tr>
<tr>
<td>Johnson</td>
<td>527-575</td>
<td>neighborhood features or residency as predictors or treatments;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>outcomes that were measures of achievement, psychosocial dispositions toward education, matriculation, or learning behaviors;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>contemporary U.S. populations sampled circa 1960 or later; and</td>
</tr>
<tr>
<td></td>
<td></td>
<td>estimates of neighborhood influences for African Americans.</td>
</tr>
<tr>
<td>Bohnert et al.</td>
<td>576-610</td>
<td>utilized at least one of the dimensions of involvement (i.e., breadth, intensity, duration, engagement).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>both quantitative and qualitative in nature and utilized survey, experience sampling, time-use diaries, and observational methods for data collection.</td>
</tr>
<tr>
<td>Lovett</td>
<td>611-638</td>
<td>none</td>
</tr>
</tbody>
</table>