METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Original Article

Check for updates

Extending the Reach of the Common Cause Design Using Meta-Analytic Methods: Applications and Issues

Christopher G. Thompson [1], William H. Yeaton [2], Gertrudes Velasquez [2], Kitchka Petrova [2], Betsy J. Becker [2]

[1] *Department of Educational Psychology, Texas A&M University, College Station, TX, USA.* **[2]** *Department of Educational Psychology & Learning Systems, Florida State University, Tallahassee, FL, USA.*

Methodology, 2024, Vol. 20(4), 238–264, https://doi.org/10.5964/meth.10591

Received: 2022-11-04 • **Accepted:** 2024-10-14 • **Published (VoR):** 2024-12-23

Handling Editor: Isabel Benítez, University of Granada, Granada, Spain

Corresponding Author: Christopher G. Thompson, Department of Educational Psychology, Texas A&M University, College Station, TX, 77843, USA. E-mail: cgthompson@tamu.edu

Abstract

Most meta-analytic methods examine effects across a collection of primary studies. We introduce an application of meta-analytic techniques to estimate effects and homogeneity *within* a single, primary study consisting of multiple, pretest-intervention-posttest units. This novel assessment was used to validate the recently created "Common Cause" (CC) design. In each case, we established the CC design by eliminating control groups from randomized studies, thereby deconstructing each experiment. This deconstruction enabled us to compare difference-in-difference results in randomized designs with a control group to pretest-posttest differences in a CC design without a control group. Meta-analysis results of multiple OXO effects from the CC designs were compared to meta-analytic effects of multiple randomized studies. This within-study-comparison logic and associated analyses produced consistent similarity between CC and validating-study results when directions of findings and patterns of statistical significance were considered. We provide plausible explanations for varying CC effect-size estimates, describe strengths and limitations, and address future research directions.

Keywords

quasi-experimental design, research methods, meta-analysis, John Stewart Mill

The Common Cause (CC) research design (Yeaton & Thompson, 2016) is a relatively new creation. In basic terms, the design is constructed by repeating a given intervention across multiple study units. The design is particularly useful when control-group data

are not available or if information on pertinent, between-units confounding variables are lacking. The logic of CC relies on the aggregation of effects from multiple OXO units wherein pre-intervention and post-intervention measures are represented as "O"s and the common intervention as an "X".

The logical foundation of the CC design was introduced more than a century and a half ago by the philosopher John Stuart Mill (Mill, 1843). While the social and medical sciences commonly use the "Method of difference" principle (where an intervention group differs from a control group in only one way – the presence of the intervention), the CC design applies Mill's "Method of agreement." This philosophical canon argues for cause via a pattern of consistent benefit (or harm) across multiple OXO units. In Xs and Os notation (Yeaton, 2019), we can write the design as

OXO (Unit 1)
•
•
•
OXO (Unit *k*).

If potential confounding variables and explanations of cause within individual OXO units are held constant across units or vary in a sufficiently random way, causal inference is enhanced.

From a conceptual perspective, the OXO design has a long history in biology, physiological psychology, chemistry, and physics. In these contexts, researchers often apply the design within well-controlled environments. Inorganic materials are often stable, and hypotheses are made regarding rates of change that are well known and for which magnitudes of change are predictable. Here, an intervention may be presented to a single or small number of units, then withdrawn and re-introduced. In the absence of such conditions, causal inference is less sound. These "conditions of use" represent invaluable surrogates for the application of CC.

Contemporary examples of CC logic are common. Large decreases in air pollution co-occurred in many large cities around the world when residents stayed at home due to the coronavirus pandemic (Chauhan & Singh, 2020). Similarly, reduced travel was accompanied by a decrease in Covid-19 cases and that pattern was repeated across U.S. states. When stay-at-home orders were lifted, recurring spikes in the prevalence of coronavirus were common.

An early example of the CC design in the *New England Journal of Medicine* (Phillips & Carstensen, 1986) demonstrated that after 38 television news programs or documentaries on the topic of suicide were introduced on different days, actual teenage suicides exhibited pre-to-post increases consistent with those staggered introductions. Kazdin (1981) applied the CC design in a psychotherapy context where therapists aimed to make causal

claims across their caseloads of clients. Despite the presence of potentially confounding variables, replication of beneficial interventions for individual psychotherapeutic cases allowed Kazdin to cautiously claim success for a treatment regimen (the "X" in each OXO).

Internal Validity Threats in the CC Design

The building block of the CC design is a pretest-intervention-posttest unit. Ultimately, the veridicality of the CC design fundamentally depends upon the quality of OXO units. Readers might not be immediately convinced that a set of individually weak pieces of evidence is inferentially sound, in the aggregate. Indeed, as Cochran (1972, p. 128) noted, "Single-group studies are so weak logically that they should be avoided whenever possible." In contrast, Reichardt (2019) argues that "…the pretest-posttest design will not be biased by threats to internal validity in all research settings…" (p. 110). If validity threats are not consistently present across OXO units in the CC design and the magnitudes and directions of bias are sufficiently variable, the potentially biasing impacts of threats may "cancel out."

While a single OXO design is often inferentially weak, the CC design aggregates multiple OXO units. A critical element of the "Method of agreement" argues that it is the "circumstance in common" from multiple OXOs that enhances causal inference. Shadish et al. (2002) note several weaknesses germane to causal inference based on a single OXO unit (e.g., history, maturation, regression, instrumentation, and testing). Fortunately, the practical application of a CC design using multiple OXOs often mitigates these potential inferential flaws.

In the CC design, history is unlikely to threaten causal inference, since it is rare that each intervention occurs at the same time (i.e., no external event will coincide with treatment in each case). If interventions are not systematically introduced in response to trending outcomes or to a temporary problem, the chances are also minimal that the existence and direction of maturation effects would be the same for all individuals, or that any regression influence would be consistently present and in the same direction for each OXO unit. To ameliorate instrumentation risk, researchers can institute protections to ensure the mechanism of measurement does not change from pre- to post-test. Testing (or "re-testing") is not problematic when studies use naturally occurring, unobtrusive outcome measures rather than formally administered tests. Across OXO units, potential confounding variables may occur unsystematically, thereby reducing the risk of selection bias. Thus, when internal validity threats germane to individual OXO units or to multiple units are not viable, a composite of staggered OXO strands of evidence can reasonably enhance causal inference.

Validating the Common-Cause Design Using an Embedded-Study Approach

A primary purpose of the current study is to establish the degree to which effects established by the CC design can be comparable to those of a set of validating studies that have used a more rigorous research design. One approach to validation would be to compare estimates in a set of research articles using the CC design to estimates from a set of validating studies asking the same research question. However, such independent sets of studies may differ in ways that weaken their comparison. Instead, we deconstructed the design in a single set of validation studies to create a CC design. In this way, between-study confounds were avoided; estimates in CC and estimates in the validating study were made in the context of the same study conditions: Both designs used the same pretest and posttest measures, the same settings, and the same interventions, implemented in the same way.

This logic is not new. Researchers have previously utilized this within-study comparison (WSC) approach (Cook et al., 2008) to validate weaker quasi-experiments, typically by comparing randomized-control-trial (RCT) results with those found in strong quasi-experiments: the nonrandomized control group (Shadish et al., 2008), regression discontinuity (Shadish et al., 2011), and controlled interrupted time series designs (CITS; Fretheim et al., 2013; St. Clair et al., 2014). Less frequently, stronger quasi-experiments have been used to validate results in other, non-RCT designs. Somers et al. (2013) utilized regression discontinuity to validate the CITS design, while Kowalski et al. (2017) used a CITS to validate the weaker regression-point-displacement design.

WSC precedents exist for the embedded validation approach (e.g., Wong & Steiner, 2018). In our research study, two examples clearly illustrate the design-in-design approach, each resembling the deconstruction technique we used below to establish the CC design and illustrate the approach. A *New England Journal of Medicine* article (Jena et al., 2017) demonstrated that delayed hospital arrival due to traffic delays accompanying marathons in 11 cities over 11 years accounted for a 3% to 4% increase in 30-day mortality for acute myocardial infarction and cardiac arrest. The controlled-time-series design added five pretest and posttest measures for each study unit (on the same day of the week, for five weeks immediately prior to and after each marathon). Control-group data were drawn from geographic areas with ZIP codes adjacent to intervention cities, and several surrogate analyses were implemented to eliminate internal validity threats.

This strong quasi-experimental design can be represented in Xs and Os notation:

NR OOOO**OXO**OOOOO
NR OOOOO OOOOOO

As shown in bold in the first line, the CC design can be established from the 121 realizations in the CITS design by creating 121 (11 marathons x 11 years) OXOs. If the aggregation of 121 OXO differences showed mortality increases consistent with estimates from the validating study in the above CITS design (which included multiple pretests and posttests with control groups), then the CC design would have been validated.

Thomas Cook and his colleagues (Chaplin et al., 2018) used a WSC decomposition strategy to validate the regression discontinuity (RD) design. They eliminated a portion of pretest and posttest treatment and control-group results in randomized studies on either side of an arbitrary cut-point to create RD designs for five of their 15 within-study comparisons. In our CC case, RCT control-group results were eliminated to create multiple OXO units. In Chaplin et al. and in the CC case, meta-analysis methods were used to aggregate and to identify possible biasing results in quasi-experiments.

In the Chaplin et al. validation study, no efforts were taken to adjust for potential RD bias. For validation of CC, the mechanism for treatment assignment for an OXO unit from each RCT was fully known and based on chance. As in Chaplin et al., the presence of bias for CC is assessed via empirical determination. As previously noted, confounds that mediated the causal relationship might be sufficiently random across CC units and, on average, "cancel out," different confounds might produce similar effects, or confounds might be similar across units and produce bias. In our validation, we did not analyze the degree to which these three specific possibilities were present.

In the context of causal modeling, three assumptions—unconfoundedness, positivity, and consistency—are required to identify average treatment effects (Keller, Wong, Park, et al., 2024). In each RCT validating study, unconfoundedness was satisfied, as potentially confounding variables in treatment and control conditions were probabilistically equivalent by random assignment. Positivity was satisfied because the probabilities of placement of individuals into a treatment and control group will be equal to one half by expectation when randomization is successful. Consistency was satisfied given a single version of a treatment and control group in each RCT. In addition, we assumed no interference between treatment and control participants as implementation of study conditions was under researchers' control.

In the multiple OXO portions of the CC design, a single treatment group was chosen from each RCT used in the meta-analysis. Unconfoundedness was not an issue as only one treatment group was taken from each RCT. Positivity was satisfied as the chance of assignment to the treatment group in each RCT used to establish an OXO unit, in theory, had a probability equal to one half. Consistency was not relevant as, again, a single OXO condition from an RCT was used to create each Common Cause unit.

Regarding the quality of our causal-model assumptions, unconfoundedness could be assessed by an empirical assessment of the pretest equivalence of potential confounders in each RCT, were all such confounders reported. As noted above, positivity was ensured since the probability of placement of individuals in each study condition was theoretical-

ly one half. Finally, an assessment of consistency would require monitoring of treatment implementation as well as communication between treatment and control participants in each RCT. Unfortunately, few if any of our studies reported this issue.

Moving Beyond Simulated Data to Validate the CC Design

In the first CC paper (Yeaton & Thompson, 2016), ANOVA-based statistical procedures were used to analyze simulated data. The authors mimicked a variety of cases and accompanying patterns of results that could likely have occurred in CC designs. Effect-size measures were not calculated. A primary purpose of the current paper is to validate the CC design using empirical data from previously reported research. We aimed to assess the consistency of effect-size estimates by comparing results in validating and CC studies. To the extent that CC-design conclusions using multiple OXO units agree with conclusions in a set of validating studies implementing a variety of stronger designs, the CC design can subsequently be implemented "by itself" and be expected to lead to similar effect estimates, under comparable sets of conditions. If the CC and validating original study results were not in agreement, an important limiting condition would be established.

A variety of research questions and outcomes were addressed in this study. We empirically tested CC in two cases: a physical intervention named "cupping" meant to ameliorate pain; and tutoring aimed to increase Scholastic Aptitude Test (SAT) performance. These cases represented disparate fields including education, health, and psychology. In each case, subsets of effects were also analyzed.

Our current study included two validation cases in which primary data in both sets of validating studies were meta-analyses of RCTs. In each case, the statistical assessment of the OXO design was based upon meta-analytic data drawn from the original RCT studies. Evidence for the validity of the CC design was established using the multiple, primary RCT studies in each meta-analysis. In both of our meta-analysis-based cases, a difference-in-differences (DID) approach had been implemented in the validating study. To create each CC design, we deconstructed the control-group design to instead aggregate only the OXO (uncontrolled intervention) portion of each group-comparison study from the validating meta-analysis. We also assessed effect-size homogeneity across OXO and validating RCT units. In addition to assessing comparability of the overall effect sizes in the CC and the validating study, we examined the patterns of statistical significance in both the CC and validating study (VS).

Precedents for a Multiple Evidentiary Approach

Long-standing precedents support the use of multiple strands of evidence to assess claims. The U.S. justice system strives to use numerous kinds of arguments to establish guilt or innocence (e.g., motive, opportunity, and history of similar crimes). In epidemiol-

ogy, researchers "play detective," attempting to find the common element that foretells the presence of a disease but is probabilistically absent when the disease is absent. This evidentiary logic is familiar to applied researchers, sometimes falling under the rubric of "pattern matching" (Shadish et al., 1986). It asks: 1) 'Is there a consistent pattern to the findings?' and, 2) 'Does that observed pattern match the predictions of the causal claim?'

When a single, well-done experimental study is not feasible, observational study researchers have been encouraged to "make your theories elaborate" to more convincingly establish cause (e.g., Cook, 2015). Rosenbaum (2015) cites Haack's (1995) use of the crossword puzzle as a logical analogue for tying together pieces of evidence both across and within studies:

> How reasonable one's confidence is that a certain entry in a crossword is correct depends on: how much support is given to this entry by the clue and any intersecting entries that have already been filled in; how reasonable, independently of the entry in question, one's confidence is that those other already filled-in entries are correct; and how many of the intersecting entries have been filled in (p. 207).

Rosenbaum (2015) further elaborates the crossword puzzle analogy in a way that mirrors its use with the CC design—when the individual OXOs are non-intersecting:

> In a crossword puzzle, entries need not intersect to provide mutual support. If 2 down meets both 4 across and 6 across, then an entry in 6 across may support the entry in 2 down, and the entry in 2 down may support the entry in 4 across, so the entry in 6 across supports the entry in 4 across even though 6 across and 4 across do not intersect (p. 208).

Comparability of CC- and Validating-Study Results

In their original CC design paper, Yeaton and Thompson (2016) provided an explicit basis upon which to judge consistency among OXO units—a statistically non-significant interaction test. Another standard commonly used to judge consistency uses the pattern of chance findings as the relevant criterion. As Shadish and Cook (2009) suggest:

> Given such a pattern matching logic, statistical analyses are required that test the overall fit of all the hypothesis tests, not just the difference between adjacent means as in the simple designs. But such tests are not as well developed as those for testing the difference among a small number of means. It may be that testing effects in pattern matching designs requires an approach more resembling meta-analysis, such as combined probability tests. This is a topic needing considerable attention (p. 623).

To further clarify, what if 15 of 20 individual pretest-posttest OXO differences were in a positive direction? Furthermore, what if 12 of these 15 positive differences were statistically significant? Is this level of consistency sufficient to affirm a casual claim? (However, see Hedges & Olkin (1980) on the weaknesses of this simplistic type of approach). What about the magnitude of the difference between the aggregate estimate in the set of OXO results and the estimate in the validating study? Could some minimal "closeness" criterion establish comparability between effect estimates in the two designs?

Steiner and Wong (2018) addressed precisely these kinds of questions in judging what they term "correspondence" between study results. These authors distinguished between two kinds of research questions. The first touches upon the "policy issues" being addressed. Practically, they asked if policy makers would "draw the same conclusions" from both kinds of studies. In this context, they regarded "direction and magnitude of effects as well as statistical significance patterns" to be of paramount importance. Second, for methodological questions, they focused upon "distance-based correspondence measures" to judge bias and argued for direct statistical tests of difference or equivalence.

We relied upon both conceptualizations (direction and distance) when assessing the comparability of CC results (an aggregate of OXO differences) and results of the validating study (which used pre-post differences in both the intervention and control groups). When choosing a between-designs difference threshold to gauge comparability, we were initially guided by Steiner and Wong's (2016) decision that "…a relatively large tolerance threshold of at least 0.30 *SD* was needed for establishing equivalence in unbiased benchmark and non-experimental estimates" (p. 27). However, it is more common for methodological researchers to opt for smaller criteria to attain adequate "closeness" (e.g., the 0.10 *SD* distance used by Chaplin et al. (2018) and advocated by Kruschke (2018), often referred to as a ROPE or Region of Practical Equivalence). We report *SD* differences for RCT- and CC-based estimates as empirically established in both study cases.

Method

In the original CC publication, Yeaton and Thompson (2016) assessed statistical significance of OXO estimates but did not calculate effect-size estimates. As it was natural to aggregate individual effect estimates within a CC study with multiple OXO units, for our research, we looked to relevant meta-analytic statistical techniques where syntheses of such between-study OXO differences are relatively common. Fortuitously, well-documented meta-analytic methods exist for assessing magnitude and variability of multiple OXO units in primary studies (e.g., Becker, 1988). That is, we were able to apply meta-analytic statistical formulae from sets of primary studies for our CC cases in which OXO units and RCT units fell *within a single study*. This novel application of meta-analysis

was responsive to Shadish and Cook's plea for appropriate methods to establish pattern matching and provided a sound framework to assess quality in the CC design.

Overall Effect Sizes and Effect-Size Consistency: Statistical Considerations

As noted above, in our meta-analytic-based validation cases, to create an OXO result we deconstructed the original treatment/control-group design and effect size in each primary study and then computed the mean and variance of the new effect sizes for a series of OXO units.

To denote a generic effect size to allow for description of our analyses, we use the sample effect size, T_i, and its variance estimate, v_i. For $i = 1, ..., k$ (with the number of studies denoted as k), the random-effects weighted estimate of each overall effect was computed as

$$\hat{\mu} = \frac{\sum_{i=1}^{k} \frac{T_i}{v_i + \hat{\tau}^2}}{\sum_{i=1}^{k} \frac{1}{v_i + \hat{\tau}^2}},$$

and the estimated variance of the overall effect was

$$V(\hat{\mu}) = \frac{1}{\sum_{i=1}^{k} \frac{1}{v_i + \hat{\tau}^2}}.$$

Here, $\hat{\tau}^2$ is an estimate of the between-studies (e.g., between OXOs) variance (for example, Chapter 12 of Borenstein et al., 2009). In the CC context, note that k represents the number of OXO (or RCT) effects.

Another integral component of any meta-analysis is the evaluation of effect-size homogeneity. Because effect sizes can be estimated from individual OXO results using meta-analytic techniques, it is natural to consider meta-analytic methods to assess homogeneity in the CC and validating-design effects. We use two measures of homogeneity to evaluate consistency of results: the Q statistic (e.g., Hedges, 1982) and the I^2 index (e.g., Higgins & Thompson, 2002; Higgins et al., 2003).

First, to assess similarity of the OXO effects, Q can be used to test the null hypothesis of homogeneity. Essentially, we ask whether OXO results are consistent or comparable across all OXO units within the CC study. We compute Q as the weighted variance

$$Q = \sum_{i=1}^{k} v_i^{-1} \left(T_i - \left(\sum_{i=1}^{k} v_i^{-1} T_i \right) \left(\sum_{i=1}^{k} v_i^{-1} \right)^{-1} \right)^2.$$

The same index is computed for the validating RCT effects.

We interpret Q as a ratio of observed, between-studies variation to within-study error using the inverse of v_i (Borenstein et al., 2009). If Q is larger than the critical value of a chi-square distribution with $k-1$ degrees of freedom (at a given α such as .05 or .01), we reject the null hypothesis of effect-size homogeneity. Large Q values indicate that our results do not all agree but do not imply that each OXO (or RCT) result is unique.

To provide potentially corroborative evidence, either for or against OXO consistency, we use a second measure of homogeneity, I^2, which estimates the percent of effect-size variability not explained by always-present sampling error. We compute I^2 as

$$I^2 = \left(\frac{Q-k+1}{Q}\right) \times 100\% .$$

In the CC context, the larger the I^2 value (which ranges from 0% to 100%), the larger the degree of heterogeneity among OXO results. We also compute I^2 for the validating RCT data. Higgins et al. (2003) proposed rules of thumb where $I^2 \leq 25\%$ represents low heterogeneity, I^2 values near 50% reflect moderate heterogeneity, and larger values strong heterogeneity.

To summarize, we utilized meta-analytic methods and within-study-comparison logic to provide comparisons of RCT results and CC results decomposed from the original RCTs. We used these comparisons as a basis to investigate potential bias and to validate outcomes for CC units. Below, we note our approach for estimating average OXO effects and for assessing the heterogeneity of the results of the RCTs and CCs. We then present evidence from our two validating cases.

General Method

We chose two cases that included different content domains: cupping, which relied upon health and psychological principles; and SAT coaching, which concerned test-preparation procedures for pre-college students. While we focused upon RCTs as the preferred validating design, by including NECGs in the coaching example we were able to assess the quality of a commonly used quasi-experimental design as a potential basis of validation.

Meta-analyses of the CC design were completed in three stages. First, each OXO unit was treated as a "data point" and an effect size was computed for each. The individual effect sizes (T_i) in each example corresponded to standardized pre-post mean-change scores between the two "Os" in a given OXO unit. Second, at the meta-analytic stage, statistical homogeneity of effect sizes was evaluated using Q and I^2 defined above. Third, we estimated the magnitude and direction of the overall effect based on the CC design using random-effects weighted means. The random-effects model provided an estimate of overall effect weighted by within-OXO variability and between-OXO variability. (The

Table 1

Dimensions of Validating and Common-Cause Studies

	Example Study	
Dimension	**[a]Cupping MA**	**[b]SAT Coaching MA**
Research Question	Compare cupping to no cupping or to alternative	Compare SAT coaching to non-SAT coaching
Independent Variable	Cups placed on skin	Coaching
Dependent Variable	Self-reported pain	Verbal and Math scores
Validating Study: Method of Analysis	MA of effects by treatment area and cupping type	MA of effects by primary study design and content area
Common-Cause Study: Method of Analysis	OXO aggregates	OXO aggregates

Note. MA = Meta-analysis.
[a] Velasquez & Becker (2019). [b] Becker (1990).

latter variability component, denoted $\widehat{\tau}^2$ above, was estimated using restricted maximum likelihood).

In each case, we compared estimates from the CC design to meta-analytic estimates from the original validating studies. A particular strength of our analyses is that the OXO and validating-study effect sizes are paired, as both are drawn from the same primary-study samples. This ensures comparability of the two designs because the effects draw on the same populations measured using the same measurement instruments. To the extent internal validity threats such as history and maturation were present, both effects would be similarly impacted. Table 1 contains descriptions of the important dimensions of the validating and CC studies in our two cases.

Methods for Individual Cases

Case 1: Cupping Meta-Analysis

Cupping, also known as "myofascial decomposition," is a pain-reduction technique studied primarily in the complementary and alternative medicine literature. The full report of the synthesis of cupping studies (Velasquez & Becker, 2019) is in the process of submission for publication. To gather studies of the effects of cupping on body pain, an electronic search was conducted for studies published between 2009 and 2016 using multiple databases: PubMed, BioMedCentral, CINAHL, ScienceDirect, Google Scholar, Dissertation and Theses (Global), and a university interlibrary loan service. The search used several keywords: "cupping" OR "myofascial decomposition" AND "back pain" OR "shoulder pain" OR "neck pain" OR "body pain." The initial search yielded 30 studies,

some of which were duplicates, qualitative systematic reviews, or studies that did not report appropriate statistical information for quantitative analysis. After further review, 16 RCTs were considered and 13 were chosen (three RCTs did not report pain intensity).

Description of Included Studies — The cupping treatment consists of the placement of small cups on the skin of participants in the region that is painful. Study participants assigned to treatment groups received either wet cupping, dry cupping, or a combination of both procedures. Study participants in control groups received some other mainstream or alternative pain therapy, medication for relieving pain, or no treatment. Mainstream or alternative therapies included application of heat packs or pads, relaxation exercise, or acupuncture. Pain medications included physician-recommended dosages of diclofenac sodium or dexibuprofen. Participants were not experiencing pain due to extreme physical trauma and were not restricted from participation based on demographic characteristics or level of physical activity.

Pain intensity, as reported in the 13 original studies, was selected as the outcome measure for the validating study and for the CC case. This continuous outcome was measured using either a Visual Analog Scale, a Numeric Rating Scale, or the Present Pain Intensity scale of the McGill Pain Questionnaire (Melzack, 1987). One primary study reported two pain-intensity measures, one for the shoulder area and another for the neck area. In this single instance, one measure was randomly selected for CC inclusion. Sample size at pretest and unstandardized pretest and posttest means and standard deviations of pain intensity were coded in each primary study. Posttreatment *SD*s of the pain-intensity variable were not reported in two studies but were calculated from reported standard errors and 95% confidence intervals. The posttreatment *SD* of the outcome variable for a third primary study was obtained from its first author.

Coding and Reliability — Two students in a graduate measurement and statistics program independently coded 68 variables in the primary studies. Percent agreement between coders on variables used for CC ranged from 90% to 100%. Coders resolved all disagreements before inclusion of agreed-upon findings in the meta-analysis.

Effect-Size Calculation — Using pain-level summary statistics from the primary studies, standardized-mean-change (SMC) effect sizes for the treatment and control groups, d_{T} and d_{C}, and their variances, $V(d_{\mathrm{T}})$ and $V(d_{\mathrm{C}})$, were calculated using formulas given in Becker (1988). The SMC for treatment, d_{T}, is the effect size (i.e., the T_i above) for the CC analysis. The SMC effect sizes are given by

$$d_{\mathrm{T}} = c(n_{\mathrm{T}} - 1)\frac{\bar{Y}_{\mathrm{T}} - \bar{X}_{\mathrm{T}}}{S_{X,\,\mathrm{T}}}, \tag{1a}$$

$$d_C = c(n_C - 1)\frac{\bar{Y}_C - \bar{X}_C}{S_{X,\,C}}, \tag{1b}$$

where $\bar{X}_T$, $\bar{Y}_T$, n_T, and $S_{X,\,T}$ are the pretest and posttest means, sample size, and pretest standard deviation for the cupping group and $\bar{X}_C$, $\bar{Y}_C$, n_C, and $S_{X,\,C}$ are analogues for the control (non-cupping) group. Bias-correction factors[1], $c(n_T - 1)$ and $c(n_C - 1)$, were applied to the respective SMC effect-size calculations.

The sample variances of the SMC effect sizes are

$$V(d_T) = \frac{2\left(1 - r_{(\text{pre, post})_T}\right)}{n_T} + \frac{d_T^2}{2n_T}, \tag{2a}$$

$$V(d_C) = \frac{2\left(1 - r_{(\text{pre, post})_C}\right)}{n_C} + \frac{d_C^2}{2n_C}, \tag{2b}$$

where $r_{(\text{pre, post})_T}$ and $r_{(\text{pre,post})_C}$ are sample pretest-posttest correlations for the cupping group and control group.

The difference in SMC effect sizes

$$d = d_T - d_C, \tag{3}$$

was used in the original meta-analysis of RCTs as T_i. It represents the effectiveness of the cupping treatment in treating body pain beyond what would be expected from using mainstream or alternative (control) pain therapies. The sample variance of d, denoted as $V(d)$, is given by

$$V(d) = V(d_T) + V(d_C). \tag{4}$$

Case 2: Meta-Analysis of SAT Coaching

Description of Included Studies — For our second case, we reanalyzed data from a meta-analysis on the effectiveness of SAT coaching on math and verbal SAT outcomes (Becker, 1990). The SAT (formerly Scholastic Aptitude Test) is a widely used college entrance examination. These analyses also used standardized, DID effect-size measures (Becker, 1988) between two time points (pretest and posttest), as defined in Equations 1a through 4. Becker (1990) included four designs: uncontrolled designs, non-equivalent group comparisons (NEGCs), and randomized and matched trials. For this example, for validation purposes, we used only studies with non-equivalent comparisons and randomized-study designs. Effect sizes for OXO intervention groups from primary studies in the

1) Hedges' correction for bias, $c(m) = 1 - \frac{3}{4m-1}$ with m degrees of freedom, serves as the bias-correction factor in the set of cupping studies.

meta-analysis were used in our CC study (17 effects for math outcomes and 27 effects for verbal outcomes). For the validating study, we used the DID effect size d from Equation 3 for the same studies that provided the OXO effects. As for the cupping example, d was a difference between the standardized-mean-change scores reported for each coached (i.e., OXO) and control group.

Effect Sizes and Reliability — In this instance, SMC effect sizes d_T and d_C as well as the effect size d had previously been computed and reported by Becker (1990, p. 410). However, to apply meta-analytic methods associated with the CC design, we needed relevant sample variances for effect size d_T, which had not been reported in Becker (1990). Using Equation 13 in Becker (1988), a reported pre-posttest correlation of .88 (for both math and verbal outcomes) and information in Appendix A of Becker (1990), we computed the separate sample variances $V(d_T)$ and $V(d_C)$. Because acceptable inter-coder reliability had been reported by Becker (1990), reliability was not recalculated for SAT-coaching effect-size measures. As above, Table 1 contains a description of important dimensions of validating and CC studies of SAT coaching.

Effect-Size Calculation for SAT-Coaching Effects — For the SAT-coaching example, we again used a DID estimator for the validating study that was different than the weighted mean, change-score estimator used for CC. The process of computing CC and validating-study results was as follows.

First, we extracted two or four standardized mean-gain scores (see Equations 1a and 1b) for each individual primary study, in both RCT and nonequivalent-comparison-groups (NECG) designs. Four effects arose because two groups (coached and uncoached) took either the Math SAT, the Verbal SAT, or both. These effect sizes were taken from Becker (1990, p. 410).

The standardized mean-gain scores for the coached groups served as the CC effect sizes. Next, for the validating study, we computed using Equation 3 the differences d^M and d^V between the standardized mean-gain scores of the treatment and control groups for Math and Verbal SATs, respectively. In total, four outcome combinations were created (two disciplines by two study designs, RCT and NEGC).

Results

In this section, we report results from each case, cupping and SAT coaching. Meta-analytic CC findings consist of an aggregate measure of central tendency ($\hat{\mu}$), its standard error (SE), and homogeneity indices (Q and I^2). For all findings, we assessed statistical significance. This approach effectively treats the intervention results from the RCTs in both examples as if a series of OXO designs had been observed.

Cupping Meta-Analysis

The cupping data consisted of 26 effect sizes for pain intensity, with two (d_T and d) for each of the 13 primary studies. The effect sizes used in Becker (1988) require a pre-post correlation estimate for variance computation; this value was rarely reported in the cupping studies. We assessed the sensitivity of pre-post correlation choice by computing the meta-analyses twice, using a lower bound (.49) and upper bound (.69) of the mean of the reported correlations. Results using the lower and upper bounds on the pre-post correlation were robust to the choice of pre-post correlation (see Table 2), thus we discuss only results based on the lower-bound correlation (i.e., the more conservative choice).

Using the CC design and effect size d_T with the lower pre-post correlation, our weighted random-effects mean was $\widehat{\mu}_{CC} = -1.73 \quad (SE = 0.27, \quad p < .001)$, indicating a substantial reduction in pain intensity due to cupping. We found statistically significant effect-size heterogeneity, $Q_{CC}(12) = 135.80$, $p < .001$, which was also reflected by $I^2_{CC} = 92\%$.

Parallel analyses of the DID effect sizes d from the treatment-control studies produced a mean of $\widehat{\mu}_{VS}$ = -1.57 (SE = 0.32, p < .001) and VS effect-size heterogeneity was notable, with $Q_{VS}(12) = 204.60 \quad (p < .001)$ and $I^2_{VS} = 96\%$.

In summary, the results in the CC design ($\widehat{\mu}_{CC}$ = -1.73) and in the validating study ($\widehat{\mu}_{VS} = -1.57$) both revealed large reductions in pain. The 0.16 SD between-designs difference was well within the 0.30 correspondence standard suggested by Steiner and Wong (2016).

In addition to estimates based on all studies, we calculated separate results for dry and wet cupping and by cupping site, either back or neck. In three of these four subsets, the CC design reflected a larger effect than the validating study (the average overestimate was about 0.11 SD). All four CC and validating subsets reported in Table 2 showed statistically significant benefits that consistently favored cupping.

A direct comparison of these two effect-size estimators can be challenging as they represent two different causal estimands: one is a mean change score across groups (CC) and the other is a mean difference-in-differences between two groups, for a set of treatment and control groups. However, Zelinsky and Shadish (2018) conducted a meta-analysis of both change measures and between-groups effect sizes in the context of single-case designs, much like our data. Conditions required for equivalence of such indices, including normality of scores, absence of a time trend, and standardization based on total between and within variation have been outlined by Shadish and others (2014). While some of these conditions do not apply to our OXO data structure (e.g., time trends within each O cannot be observed with only one measure at pre and post), we consider the pretest as analogous to a control and the posttest observation as parallel to a treatment group, and our OXO mean-change measure meets the other conditions outlined by Shadish et al. (2014).

Methodology
2024, Vol. 20(4), 238–264
https://doi.org/10.5964/meth.10591

Table 2

Cupping: Common Cause and Validating-Study Results

	[a]Common-Cause OXO Results				Validating-Study RCT Results				Comparison	
Dataset	**RE Mean**	***SE***	**Direction**	**Sig?**	**Effect**	***SE***	**Direction**	**Sig?**	**CC – VS Difference**	**Sig?**
Full	-1.73	0.27	Negative	Yes	-1.57	0.32	Negative	Yes	-0.16	No
Treatment Area										
Back	-2.12	0.31	Negative	Yes	-2.07	0.50	Negative	Yes	-0.05	No
Neck	-1.10	0.38	Negative	Yes	-0.94	0.16	Negative	Yes	-0.16	No
Cupping Type										
Dry	-1.91	0.35	Negative	Yes	-1.92	0.51	Negative	Yes	0.01	No
Wet	-1.44	0.44	Negative	Yes	-1.21	0.34	Negative	Yes	-0.23	No

Note. RE = Random-Effects; *SE* = Standard Error; Sig? = Statistical significance of mean effect at $\alpha = .05$ level. A pre-post correlation of $r = .49$ was used to compute the variance of d^T.

[a] Based on data from Velasquez & Becker (2019).

As both of our estimates can be interpreted as *d*-type evidence, they share some important characteristics. Both effects show large reductions due to treatment in absolute terms, given the standardized scale of the metric. Also, both full and subset effects were negative (reflecting pain reduction) and were statistically significantly different from the null value of zero effect. For the random-effects CC model, the full dataset difference between the two approaches $\widehat{\mu}_{CC} - \widehat{\mu}_{VS} = 0.16$, is unlikely to be judged clinically significant, falling below a 0.30 *SD* comparison threshold. In addition, both sets show heterogeneity. While this implies that individual effects are expected to vary around the two means $\widehat{\mu}_{CC}$ and $\widehat{\mu}_{VS}$, all observed values of d_T and d were negative, another indicator of the similarity between the CC and validating studies.

SAT Coaching Meta-analysis

Results for the SAT coaching data (see Table 3) are reported by subject area (Math or Verbal) and study design (NECG or RCT). For our validating study, three of the four subgroups results had statistically significant means, two at the .01 level and one at the .05 level. All four RCT-based validations of CC showed statistically significant means. For both validating data and CC results, larger effects were found for math than for verbal outcomes.

In all four CC-based analyses of the SAT-coaching data, we found evidence for statistically significant change due to coaching, with positive weighted means. All measures of coaching effectiveness in validating-study analyses were also positive and favored coaching). All four CC results also showed statistically significant effect-size heterogeneity, with all Q tests being large, and with corresponding large I^2_{CC} values (92%, 92%, 95%, and 74%, respectively).

When one compares CC and validating-study (VS) results, decisions based on statistical significance of the means matched for three of the four corresponding pairs of cells for the CC and validating studies in Table 3. Consistency in sign also holds: all mean coaching effects were positive, whether reflecting only change (CC) or change due to coaching beyond that expected for control participants in the validating study.

Effect magnitudes were generally comparable across designs, as reflected by the small differences between designs for the CC-VS entries in Table 3. Three of the four CC-VS differences were non-significant and less than 0.20 *SD*s, thus also below a 0.30 *SD* comparability standard. The CC effects were larger (often by a large degree) than the corresponding effects in the validating study. Also, NECG effects were larger than RCT effects for CC data but comparable for validating-study data, though we did not statistically compare these values. The impact of coaching on math and verbal outcomes was generally similar across designs; however, math effects were always slightly larger than verbal effects, likely reflecting ease of coaching for specific math skills.

Methodology
2024, Vol. 20(4), 238–264
https://doi.org/10.5964/meth.10591

Table 3

SAT Coaching: Common-Cause and Validating-Study Results by Subject and Design

		[a]Common-Cause OXO Results			Validating-Study RCT Results				Comparison	
Subject	**Design**	**RE Mean**	***SE***	**Sig?**	***k***	**RE Mean**	***SE***	**Sig?**	**CC – VS Difference**	**Sig?**
Math	NECG	0.56	0.10	Yes	11	0.18	0.08	Yes	0.38	Yes
Math	RCT	0.36	0.10	Yes	6	0.19	0.04	Yes	0.17	No
Verbal	NECG	0.23	0.08	Yes	13	0.08	0.05	No	0.15	No
Verbal	RCT	0.18	0.04	Yes	14	0.11	0.03	Yes	0.07	No

Note. RE = Random-Effects; *SE* = Standard Error; Sig? = Statistical significance of mean difference at α = .05 level.

[a] Based on Becker (1990).

Discussion

Conclusions based on the comparability of the CCs and validating RCT studies depended on the criteria used to determine whether the two sets of results were similar. When the direction of effect and pattern of statistical significance were used as comparison criteria, more general, policy-level conclusions exhibited a high degree of agreement. Given either CC or validating-study results to inform policy, decision makers' conclusions to implement these interventions would likely have been the same based on statistical significance from each of our validating studies. Even when different research designs (RCT and NECG) were used *within* the SAT coaching case, CC results generally replicated those in the validating study.

Comparability also depended upon the choice of threshold. When a difference threshold of 0.30 SD was used to gauge comparability, CC and validating-study results were nearly always comparable. If the threshold were set to 0.10 SD, CC and validating-study effect estimates would not be judged as comparable. Empirically, differences between validating and CC study results were consistently around 0.15 SDs, and study results were typically larger in the CC design.

For both cupping and SAT coaching, heterogeneity for RCTs vs. CC estimates was statistically significant. As noted above, while other measures of effect heterogeneity exist, these results suggest that caution should be taken regarding external validity inferences made from a single CC study result. Though average bias as reflected by the CC versus VS differences was modest, an individual CC study result may differ considerably from its respective validating-study result.

Reasons for Lack of Comparability of CC and Validating-Study Results

In each of our cases, the CC design was created by a deconstruction of the validating-study design. Using this embedded approach allowed us to substantially reduce the number of potential reasons for differences between the validating study and the CC design (different participants, different settings, different outcome measures, and different intervention operationalizations, as noted above). However, causal estimands were confounded with study type (Wong et al., 2019); the CC and the validating study used different methods of calculating effects.

The method by which DID measures were calculated in the validating studies is a critical component in the creation of an effect estimate, as mean DID values were compared to the corresponding mean CC estimates. Thus, any discrepancy between CC results (from the aggregate of OXO units) and the DID estimate (from the validating study) fundamentally depended upon the amount of change found within control units.

The smallest difference between a validating study and a CC study occurs when the OXO portion of the validating study is minimally adjusted, that is, where the pretest-

posttest difference was small in the control groups. Such minimal adjustments would likely occur for no-treatment control groups, wait-list controls, or control groups lacking elements that would substantially change behavior (these three kinds of controls are listed in Table 2 of Becker, 1990, p. 384). In line with this expectation, we found that the effects for both RCT and NECG designs used in the SAT validating study consistently had values considerably smaller than those in CC analyses (where no pre-post control group difference was subtracted).

In contrast, maximal differences between CC and the validating study will plausibly occur when the control group is associated with substantial and consistent pretest-posttest changes that will be "adjusted" using the DID approach (that is, when control-group changes are large). In this instance, the OXO (CC – VS Difference d_T) components in the validating study will be substantially balanced by the control-group d_C in the DID. In our cases, OXO results from CC designs generally estimated larger effects, probably due to the presence of active alternative-treatment controls (e.g., medications in the case of the cupping studies).

In the intermediate case, moderate differences between the CC and the validating study can occur when the size of the pretest-posttest control-group difference is inconsistent or uniformly moderate in size. For example, if control-group changes are large in many units (leading to a large adjustment) and small or negative in other units (leading to a relatively small adjustment), the overestimate in the CC design across all OXO units would be modest in size. When more prevalent small control-group changes overwhelm less common large pre-post control-group changes, overestimation in the CC design will also be more modest. A similar argument can be constructed when the difference from pre to post leads to an adjustment that adds to rather than subtracts from the OXO units in the validating study. In this case, there would be an *under*estimation of effects in CC.

Strengths and Limitations

These cases represent two topical areas (psychological pain management and standardized-test performance on college admissions tests) which were examined using meta-analytic methods within the original, validating studies. In the SAT-coaching case, we validated estimates from the CC design using both randomized and observational studies (NECG), across two subject-matter areas, math and verbal. In the cupping case, we also examined subsets of results (by site and kind of treatment). Meta-analytic methods of assessing means and gauging homogeneity transferred smoothly into the CC context. In addition, this novel application of meta-analytic methods further expanded the range of designs to which WSC methods can be meaningfully applied. Finally, the paper represents a promising prototype of methodological techniques that can be used to establish the quality of a new research design.

We have demonstrated that synthesis using weighted means and assessments of homogeneity of independent study results from *within-study* units, in contrast to com-

parison of results *between* studies, is not only a plausible but also a practical extension of traditional meta-analytic methods. The usual demands of a typical meta-analysis such as study choice, coding, and reliability were greatly diminished in the CC-validation context, as relevant results were conveniently reported for our two cases. Furthermore, statistical and estimation challenges that can arise in meta-analysis (e.g., estimation of quantities such as the between-studies variance) are applicable to the CC design as well. As this application yielded effect-size measures (rather than only p values as in Yeaton & Thompson, 2016), the resulting measures can themselves be meta-analyzed. The approach will further contribute to a systematic knowledge base when CC results are combined with standardized effect sizes from other CC studies, or suitably compared to aggregated effects using causal estimands not based on OXO results.

As noted above, the choice of ROPE bears on the degree to which the CC and validating study results will be judged as similar. Certainly, some researchers prefer an absolutely small ROPE, perhaps in the area of 0.05–0.10 SDs, to judge comparability of design estimates. The CC design results shown here do not meet that stringent requirement, though overall and subset results were consistently near 0.15 SDs. However, in some policy contexts, for example when multiple cities, states, or countries close or reopen to business, consistent change in level of desired (or undesired outcomes) will immediately inform policy.

The CC design might be best viewed as a viable first step, where possible, in informing the subsequent implementation of a randomized study. Certainly, the usual "more research is needed" caveat applies to CC. The design is new, and our conclusions were necessarily limited by the unique circumstances of the two cases we chose. Ironically, the pre-publication status of one of the validating studies enabled us to obtain more detailed information that otherwise might not have allowed us to conduct thorough statistical analyses for CC.

Future Research Directions

The application of meta-analytic methods within a single study has previously been suggested, but such implementations are still quite infrequent (Goh et al., 2016). Case et al. (2015) used meta-analytic techniques to aggregate two effects in two separate studies of power and social affiliation. Unlike the CC meta-analysis approach used here, Case and colleagues did not synthesize results of multiple units to produce effects within each of the two meta-analyzed studies.

It is natural to extend validations of CC by utilizing well done quasi-experiments as validating studies. While such contrasts are less ideal than those between CC and RCTs given the potential bias in observational studies, a recent Race-to-the Top (RTT) evaluation using the non-equivalent control-group design reflects one such potential opportunity (Petrova, 2018).

RTT was a federal program that initiated changes in educational policy at the state level aiming to impact STEM outcomes. Of 18 RTT states, pre-RTT and post-RTT science achievement data were available for 4th graders and 8th graders in 15 RTT awardee states, thus enabling application of CC. For each grade level, meta-analytic methods were applied to multiple OXO units and were used to compare CC results with regression-analysis results based on a difference-in-difference scheme for the NECG. All CC results were statistically significant, whereas the 4th grade but not the 8th grade regression-based data were significant. Petrova did not compare effect sizes between designs, as they were unavailable for a validating study.

Extending the Causal Power of CC

The CC design offers a viable alternative when control groups are lacking and when randomized studies are impossible to conduct or their evidence is not timely. Fortunately, methodological "add-ons" to buttress CC inference are straightforward. A multiple OXO approach provides repeated replication over multiple units, thereby enabling interventions to be withdrawn, reinstated, and staggered in time. To create a more inferentially sound counterfactual, a single pretest measure can be repeated to form a long baseline and to establish a pattern of pre-intervention change. Fortuitously, these more ideal conditions of feasible applicability are common. Public health (e.g., vaccination prevalence), education (e.g., standardized testing), and government (e.g., voting rates) represent but a few of the many contexts in which interventions are widespread and records routinely kept for multiple units.

Our research suggests that policy decisions from the CC design are likely to be similar to those of the validating study. In our cases the CC design was unlikely to understate benefits or to discount promising interventions. However, threats to internal validity may compromise causal inference, and substantial pretest-posttest changes in control groups that are not subtracted out of pre-post differences in OXO units can sometimes lead to CC results that overestimate treatment effects. The provision of direct adjustments (e.g., analyses using propensity scores) would be possible when original data are available for multiple OXO units of a given study in lieu of the meta-analytic approach taken here. This constitutes an area for future development of the CC design. Within the current mix of quasi-experimental designs available to applied researchers, however, the CC design represents a valuable addition to the methodological tools available to enhance judgments of cause.

Funding: The authors have no funding to report.

Acknowledgments: The authors have no additional (i.e., non-financial) support to report.

Competing Interests: The authors have declared that no competing interests exist.

References

The studies used in the cupping meta-analysis have been marked with an asterisk: these references are not specifically cited in the article.

*Akbarzadeh, M., Ghaemmaghami, M., Yazdanpanahi, Z., Zare, N., Azizi, A., & Mohagheghzadeh, A. (2014). The effect of dry cupping therapy at acupoint BL23 on the intensity of postpartum low back pain in primiparous women based on two types of questionnaires, 2012: A randomized clinical trial. *International Journal of Community Based Nursing and Midwifery, 2*(2), 112–120.

*AlBedah, A., Khalil, M., Elolemy, A., Hussein, A. A., AlQaed, M., Al Mudaiheem, A., Abutalib, R. A., Bazaid, F. M., Bafail, A. S., Essa, A., & Bakrain, M. Y. (2015). The use of wet cupping for persistent nonspecific low back pain: Randomized controlled clinical trial. *Journal of Alternative and Complementary Medicine, 21*(8), 504–508. https://doi.org/10.1089/acm.2015.0065

Becker, B. J. (1990). Coaching for the Scholastic Aptitude Test: Further synthesis and appraisal. *Review of Educational Research, 60*(3), 373–417. https://doi.org/10.3102/00346543060003373

Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical & Statistical Psychology, 41*(2), 257–278. https://doi.org/10.1111/j.2044-8317.1988.tb00901.x

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley.

Case, C. R., Conlon, K. E., & Maner, J. K. (2015). Affiliation-seeking among the powerless: Lacking power increases social affiliative motivation. *European Journal of Social Psychology, 45*(3), 378–385. https://doi.org/10.1002/ejsp.2089

Chaplin, D. D., Cook, T. D., Zurovac, J., Coopersmith, J. S., Finucane, M. M., Vollmer, L. N., & Morris, R. E. (2018). The internal and external validity of the regression discontinuity design: A meta-analysis of 15 within-study comparisons. *Journal of Policy Analysis and Management, 37*(2), 403–429. https://doi.org/10.1002/pam.22051

Chauhan, A., & Singh, R. P. (2020). Decline in $PM_{2.5}$ concentrations over major cities around the world associated with COVID-19. *Environmental Research, 187*, . Article 109634. https://doi.org/10.1016/j.envres.2020.109634

*Chi, L.-M., Lin, L.-M., Chen, C.-L., Wang, S.-F., Lai, H.-L., & Peng, T.-C. (2016). The effectiveness of cupping therapy on relieving chronic neck and shoulder pain: A randomized controlled trial. *Evidence-Based Complementary and Alternative Medicine, 2016*, Article 7358918. https://doi.org/10.1155/2016/7358918

Cochran, W. G. (1972). Observational studies. In T. A. Bancroft (Ed.), *Statistical papers in honor of George W. Snedecor*. Iowa State University Press.

Cook, T. D. (2015). The inheritance bequeathed to William G. Cochran that he willed forward and left to others to will forward again: The limits of observational studies that seek to mimic randomized experiments. *Observational Studies, 1*(1), 141–164. https://doi.org/10.1353/obs.2015.0012

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study

comparisons. *Journal of Policy Analysis and Management, 27*(4), 724–750. https://doi.org/10.1002/pam.20375

*Cramer, H., Lauche, R., Hohmann, C., Choi, K.-E., Rampp, T., Musial, F., Langhorst, J., & Dobos, G. (2011). Randomized controlled trial of pulsating cupping (pneumatic pulsation therapy) for chronic neck pain. *Forschende Komplementärmedizin/Research in Complementary Medicine, 18*(6), 327–334. https://doi.org/10.1159/000335294

*Farhadi, K., Schwebel, D. C., Saeb, M., Choubsaz, M., Mohammadi, R., & Ahmadi, A. (2009). The effectiveness of wet-cupping for nonspecific low back pain in Iran: A randomized controlled trial. *Complementary Therapies in Medicine, 17*(1), 9–15. https://doi.org/10.1016/j.ctim.2008.05.003

Fretheim, A., Soumerai, S. B., Zhang, F., Oxman, A. D., & Ross-Degnan, D. (2013). Interrupted time-series yielded an effect estimate concordant with the cluster-randomized control trial result. *Journal of Clinical Epidemiology, 66*(8), 883–887. https://doi.org/10.1016/j.jclinepi.2013.03.016

Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini-meta-analysis of your own studies: Some arguments on why and a primer on how. *Social and Personality Psychology Compass, 10*(10), 535–549. https://doi.org/10.1111/spc3.12267

Haack, S. (1995). *Evidence and inquiry*. Blackwell.

Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin, 92*(2), 490–499. https://doi.org/10.1037/0033-2909.92.2.490

Hedges, L. V., & Olkin, I. (1980). Vote counting methods in research synthesis. *Psychological Bulletin, 88*(2), 359–369. https://doi.org/10.1037/0033-2909.88.2.359

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in meta-analysis. *Statistics in Medicine, 21*(11), 1539–1558. https://doi.org/10.1002/sim.1186

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analysis. *British Medical Journal, 327*(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557

*Hong, Y., Wu, J., Wang, B., Li, H., & He, Y. (2006). The effect of moving cupping therapy on nonspecific low back pain. *Chinese Journal of Rehabilitation Medicine, 21*(3), 340–343.

Jena, A. B., Mann, N. C., Wedlund, L. N., & Olenski, A. (2017). Delays in emergency care and mortality during major U.S. marathons. *New England Journal of Medicine, 376*(15), 1441–1450. https://doi.org/10.1056/NEJMsa1614073

Kazdin, A. E. (1981). Valid instances from case studies. *Journal of Consulting and Clinical Psychology, 49*(2), 183–192. https://doi.org/10.1037/0022-006X.49.2.183

Keller, B., Wong, V. C., Park, S., Zhang, J., Sheehan, P., & Steiner, P. M. (2024). *A new four-arm within-study comparison: Design, implementation, and data.* OSF Preprints. https://doi.org/10.31219/osf.io/2gur9

*Kim, J.-I., Kim, T.-H., Lee, M.S., Kang, J.W., Kim, K.H., Choi, J.-Y., Kang, K.-W., Kim, A.-R., Shin, M.-S., Jung, M.-S., & Choi, S.-M. (2011). Evaluation of wet-cupping therapy for persistent non-specific low back pain: A randomised, waiting-list controlled, open-label, parallel-group pilot trial. *Trials, 12*, Article 146. https://doi.org/10.1186/1745-6215-12-146

*Kim, T.-H., Kang, J. W., Kim, K. H., Lee, M. H., Kim, J. E., Kim, J.-H., Lee, S., Shin, M.-S., Jung, S.-Y., Kim, A.-R., Park, H.-J., & Hong, K. E. (2012). Cupping for treating neck pain in video display terminal (VDT) users: A randomized controlled pilot trial. *Journal of Occupational Health, 54*(6), 416–426. https://doi.org/10.1539/joh.12-0133-OA

Kowalski, C., Yeaton, W. H., Kuhr, K., & Pfaff, H. (2017). Helping hospitals improve patient-centeredness: Assessing the impact of feedback following a best practices workshop. *Evaluation & the Health Professions, 40*(2), 180–202. https://doi.org/10.1177/0163278716677321

Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science, 1*(2), 270–280. https://doi.org/10.1177/2515245918771304

*Lauche, R., Materdey, S., Cramer, H., Haller, H., Stange, R., Dobos, G., & Rampp, T. (2013). Effectiveness of home-based cupping massage compared to progressive muscle relaxation in patients with chronic neck pain—A randomized controlled trial. *PLoS One, 8*(6), Article e65378. https://doi.org/10.1371/journal.pone.0065378

*Liu, B., Min, X., & Huang, C.-J. (2008). Therapeutic effect of balance cupping therapy on non-specific low back pain. *Chinese Journal of Rehabiliation Medicine Theory and Practice., 14*(6), 572–573.

Melzack, R. (1987). The short-form McGill Pain Questionnaire. *Pain, 30*(2), 191–197. https://doi.org/10.1016/0304-3959(87)91074-8

Mill, J. S. (1843). *A system of logic, ratiocinative and inductive: Being a connected view of the principles of evidence and the methods of scientific investigation.* John W. Parker.

Petrova, K. (2018). *The effects of race-to-the-top on students' science achievement* [Doctoral dissertation, Florida State University, Tallahassee].

Phillips, D. P., & Carstensen, L. L. (1986). Clustering of teenage suicides after television news stories about suicide. *New England Journal of Medicine, 315*(11), 685–689. https://doi.org/10.1056/NEJM198609113151106

Reichardt, C. S. (2019). *Quasi-experimentation: A guide to design and analysis.* Guilford Press.

Rosenbaum, P. R. (2015). Cochran's causal crossword. *Observational Studies, 1*(1), 205–211. https://doi.org/10.1353/obs.2015.0021

*Saha, F. J., Schumann, S., Cramer, H., Hohmann, C., Choi, K.-E., Rolke, R., Langhorst, J., Rampp, T., Dobos, G., & Lauche, R. (2017). The effects of cupping massage in patients with chronic neck pain—A randomized controlled trial. *Complementary Medicine Research, 24*(1), 26–32. https://doi.org/10.1159/000454872

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random to nonrandom assignment. *Journal of the American Statistical Association, 103*(484), 1334–1344. https://doi.org/10.1198/016214508000000733

Shadish, W. R., & Cook, T. D. (2009). The renaissance of field experiments in evaluating interventions. *Annual Review of Psychology, 60*, 607–629. https://doi.org/10.1146/annurev.psych.60.110707.163544

Shadish, W. R., Cook, T. D., & Campbell, T. D. (2002). *Experimental and quasi-experimental design for generalized causal inference*. Houghton Mifflin.

Shadish, W. R., Cook, T. D., & Houts, A. C. (1986). Quasi-experimentation in a critical multiplist mode. In W. M. K. Trochim (Ed.), *Advances in quasi-experimental design and analysis* (No. 31, New Directions for Program Evaluation). Jossey-Bass.

Shadish, W. R., Galindo, R., Wong, V. C., Steiner, P. M., & Cook, T. D. (2011). A randomized experiment comparing random and cutoff-based assignment. *Psychological Methods, 16*(2), 179–191. https://doi.org/10.1037/a0023345

Shadish, W. R., Hedges, L. V., Pustejovsky, J. E., Boyajian, J. G., Sullivan, K. J., Andrade, A., & Barrientos, J. L. (2014). A *d*-statistic for single-case designs that is equivalent to the usual between-groups d-statistic. *Neuropsychological Rehabilitation, 24*(3–4), 528–553. https://doi.org/10.1080/09602011.2013.819021

Somers, M.-A., Zhu, P., Jacob, R., & Bloom, H. (2013). *The validity and precision of the comparative interrupted time series design and the difference-in-difference design in educational evaluation* [MDRC Working Paper on Research Methodology]. MDRC. https://www.mdrc.org/work/publications/validity-and-precision-comparative-interrupted-time-series-design-and-difference

St. Clair, T., Cook, T. D., & Hallberg, K. (2014). Examining the internal validity and statistical precision of the comparative interrupted time series design by comparison with a randomized experiment. *American Journal of Evaluation, 35*(3), 311–327. https://doi.org/10.1177/1098214014527337

Steiner, P. M., & Wong, V. C. (2016). *Assessing correspondence between experimental and non-experimental results in within-study comparisons* (Working Paper). EdPolicyWorks. https://education.virginia.edu/documents/epw46-analysis-within-study-comparisons2016-04pdf

Steiner, P. M., & Wong, V. C. (2018). Assessing correspondence between experimental and nonexperimental estimates in within-study comparisons. *Evaluation Review, 42*(2), 214–247. https://doi.org/10.1177/0193841X18773807

*Teut, M., Ullmann, A., Ortiz, M., Rotter, G., Binting, S., Cree, M., Lotz, F., Roll, S., & Brinkhaus, B. (2018). Pulsatile dry cupping in chronic low back pain—A randomized three-armed controlled clinical trial. *BMC Complementary and Alternative Medicine, 18*, Article 115. https://doi.org/10.1186/s12906-018-2187-8

Velasquez, G., & Becker, B. J. (2019). *Effects of cupping therapy on body pain: A meta-analysis* [Unpublished manuscript]. Department of Educational Psychology & Learning Systems, Florida State University.

Wong, V. C., & Steiner, P. M. (2018). Designs of empirical evaluations of nonexperimental methods in field settings. *Evaluation Review, 42*(2), 176–213. https://doi.org/10.1177/0193841X18778918

Wong, V. C., Steiner, P. M., & Anglin, K. (2019). A causal replication framework for designing and assessing replication efforts. *Zeitschrift für Pschologie, 227*(4), 280–292. https://doi.org/10.1027/2151-2604/a000385

*Xu, M., Liu, B., Huang, C., Tang, F., Lou, Y., Liang, Z., & Liang, W. (2009). The therapeutic effects of balance cupping therapy on chronic lumbar muscle strain. *Liaoning Journal of Traditional Chinese Medicine, 36*, 1007–1008.

Yeaton, W. H. (2019). Experimental and quasi-experimental designs. In P. Brough (Ed.), *Advanced research methods for applied psychology: Design, analysis and reporting* (pp. 107–123). Routledge.

Yeaton, W. H., & Thompson, C. G. (2016). Transforming the canons of John Stuart Mill from philosophy to replicative, empirical research: The Common Cause research design. *Journal of Methods and Measurement in the Social Sciences, 7*(2), 122–143.

Zelinsky, N. A. M., & Shadish, W. R. (2018). A demonstration of how to do a meta-analysis that combines single-case designs with between-groups experiments: The effects of choice making on challenging behaviors performed by people with disabilities. *Developmental Neurorehabilitation, 21*(4), 266–278. https://doi.org/10.3109/17518423.2015.1100690