CHILD DEVELOPMENT

Child Development, xxxx 2018, Volume 00, Number 0, Pages 1–13

The title for this Special Section is **Meta-analysis and Individual Participant Data Synthesis in Child Development**, edited by Glenn I. Roisman and Marinus H. van IJzendoorn

Improving Causal Inferences in Meta-analyses of Longitudinal Studies: Spanking as an Illustration

Robert E. Larzelere
Oklahoma State University

Marjorie Lindner Gunnoe
Calvin College

Christopher J. Ferguson
Stetson University

To evaluate and improve the validity of causal inferences from meta-analyses of longitudinal studies, two adjustments for Time-1 outcome scores and a temporally backwards test are demonstrated. Causal inferences would be supported by robust results across both adjustment methods, distinct from results run backwards. A systematic strategy for evaluating potential confounds is also introduced. The methods are illustrated by assessing the impact of spanking on subsequent externalizing problems (child age: 18 months to 11 years). Significant results indicated a small risk or a small benefit of spanking, *depending on the adjustment method.* These meta-analytic methods are applicable for research on alternatives to spanking and other developmental science topics. The underlying principles can also improve causal inferences in individual studies.

Due to the paucity of randomized designs, most studies in developmental science produce causally ambiguous results. This limits the validity of causal inferences derived from meta-analyses of those studies. The objective of the current article is to introduce three techniques to evaluate and improve causal inferences from meta-analyses of longitudinal studies to compensate partially for the lack of randomized designs. The techniques include two sensitivity tests to help distinguish between actual causal effects and residual confounding (i.e., confounding that is incompletely controlled for) and a third technique to empirically evaluate the impact of potential confounds on a central meta-analytic effect size.

We demonstrate these techniques by considering the effect of parental spanking on children's externalizing behavior problems. Most parents intend spanking as a corrective action. Corrective actions epitomize the difficulty of making valid causal inferences, because corrective actions are inherently confounded with the poor prognosis of the problem being corrected (Larzelere & Cox, 2013). We decided on spanking for two reasons: number of relevant studies and the possibility of using these methods to reconcile differences in recent meta-analyses of spanking.

The fundamental problem in meta-analytic summaries of nonrandomized studies is confounding (Egger, Smith, & Schneider, 2001; Reeves, Deeks, Higgins, & Wells, 2008). An example of confounding occurs when nonrandomized groups differ on prognostic factors associated with the outcome.

This research was supported, in part, by Grant R03 HD044679 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, using data generously made available by Statistics Canada. We are grateful to B. Wade Brorsen for contributing to the derivation of the equation for the "slope method" in Appendix S2 and to William A. Larzelere for processing the correlations from the included studies and for helping to write the implementation steps in Appendix S3.

Correspondence concerning this article should be addressed to Robert Larzelere, Department of Human Development and Family Science, Oklahoma State University, 233 Human Sciences Bldg., Stillwater, OK 74078. Electronic mail may be sent to robert.larzelere@okstate.edu.

DOI: 10.1111/cdev.13097

Most children spanked more (vs. less) or prescribed *Ritalin* (vs. not) already have a poorer developmental prognosis prior to the administration of spanking or *Ritalin*. Failure to control adequately for the poorer prognosis preceding any corrective action will result in a positive association between the corrective action and subsequent maladjustment. Even after controlling statistically for imperfect covariates, the results could be based entirely on *residual confounding* (Rothman, Greenland, & Lash, 2008). That is to say, effects may be due to third variables that have not been adequately controlled. Even in studies that satisfy the first two requirements for valid causal inferences (significant association and correct temporal sequence), the difficulty of controlling perfectly for differential prognoses compromises the ability of longitudinal studies to meet the third requirement for valid causal inferences: ruling out plausible alternative explanations (Valentine & Thompson, 2013).

The persistence of a poorer developmental prognosis among those receiving corrective actions makes it difficult for longitudinal studies to demonstrate benefits associated with corrective actions (i.e., a poorer prognosis can be too much to overcome). Corrective actions shown to be associated with *greater* subsequent problems include parent–adolescent discussions about deviant behaviors or deviant peers (e.g., Deptula, Henry, & Schoeny, 2010) and helping with homework (Hill & Tyson, 2009). In studies of common disciplinary responses (Larzelere, Ferrer, Kuhn, & Danelia, 2010), psychotherapy and *Ritalin* (Larzelere, Ferrer, et al., 2010; Webster-Stratton, Reid, & Beauchaine, 2013), and child-care subsidies (Herbst & Tekin, 2014), an apparent adverse effect remained significant even after controlling for initial differences on outcomes.

Spanking as a Focus for This Illustration

Americans do a lot of spanking. Estimates of its prevalence in recent decades are as high as 94% (Straus & Stewart, 1999), although spanking has declined substantially since then (Ryan, Kalil, Ziol-Guest, & Padilla, 2016). In one older national study that oversampled disadvantaged families, parents reported spanking 3- to 5-year-olds an average of 1.9 times per week (Giles-Sims, Straus, & Sugarman, 1995). Virtually all researchers agree that this is too much spanking.

Researchers do not agree, however, on what constitutes acceptable spanking, if any. Should spanking be used only to enforce cooperation with time-out, as in the most effective clinical treatments for oppositional defiant disorder through the early 1990s (Roberts & Powers, 1990)? Or should all spanking be replaced with nonphysical discipline, as recommended by many professional societies (e.g., Society for Research in Human Development, 2013)? If so, should nonphysical alternatives include other negative consequences, such as time-out and privilege removal (Roberts & Powers, 1990) or be restricted to exclusively positive responses (Holden, Ashraf, Brannan, & Baker, 2016)? Consensus on the appropriateness of any spanking and on alternatives to replace it has been hindered by several factors including the selection bias associated with corrective disciplinary actions and a general failure to define spanking precisely.

The only scientific consensus conference on corporal punishment defined spanking as a type of corporal punishment that was "physically noninjurious; intended to modify behavior; and administered with an opened hand to the extremities or buttocks" (Friedman & Schonberg, 1996, p. 853). Existing meta-analyses have claimed to focus on this type of spanking, but they were limited by the fact that most spanking research is based on parents' responses to whether or how often they "spank or slap" their child or "use physical punishment," without excluding spanking with objects of various kinds.

Consensus has also been hindered by the different methods used in the five known meta-analyses of child outcomes of physical punishment. Three of these were based on unadjusted correlations. Gershoff's (2002a, Gershoff & Grogan-Kaylor, 2016) two meta-analyses found mean overall adverse effect sizes of d = .42 and .33 (equivalent to r = .21 and .16), respectively. The latter concluded that "there is no evidence that spanking does any good for children and all evidence points to the risk of it doing harm" (p. 465). Paolucci and Violato (2004) also used unadjusted correlations and obtained a mean adverse effect size of d = .18 (r = .09), concluding that "corporal punishment does not substantially increase the risk to youth of developing affective, cognitive, or behavioral pathologies" (p. 197).

The other two meta-analyses went beyond unadjusted correlations in different ways. Larzelere and Kuhn (2005) distinguished different types of physical punishment and examined studies that investigated both spanking and an *alternative* disciplinary response. Assuming a similar selection bias for any two disciplinary responses to misbehavior, the investigators' rationale was that the difference in effect sizes between any two disciplinary responses would approximate an unbiased causal effect better than the effect size of either disciplinary response alone. They concluded that child outcomes of physical punishment were more adverse than outcomes of alternative disciplinary responses *only* for overly severe or predominant use of physical punishment. They also isolated what they believed to be the most appropriate use of spanking, called "conditional spanking" (nonabusive, used when 2- to 6-year-olds respond defiantly to milder disciplinary responses such as time-out or reasoning). Conditional spanking was associated with *less* defiance or aggression than 10 of the 13 alternatives it had been compared with. Effect sizes for "customary spanking" (neither conditional nor predominant or severe) were similar to those for alternative disciplinary responses.

Ferguson's (2013) meta-analysis went beyond unadjusted correlations by focusing only on 45 longitudinal studies, 25 of which controlled for preexisting differences on the outcome variable. Although other meta-analyses have tested whether effect sizes differ for longitudinal correlations versus other designs (Gershoff, 2002a; Gershoff, Ansari, Purtell, & Sexton, 2016), Ferguson was the first to limit a meta-analysis to longitudinal studies and the first to incorporate the most basic controls common to longitudinal studies. Ferguson accomplished this by averaging *partial r*s (*pr*s, i.e., the association between Time-1 spanking and a Time-2 outcome that remains after partialling out Time-1 outcome scores and other covariates). Ferguson reported mean adverse effect sizes from $pr = .06$ to .11 for three outcomes, the smallest being the effect of spanking on externalizing problems for 1- to 6-year-olds. Ferguson (2013) concluded that social scientists "should take greater care not to exaggerate the magnitude and conclusiveness of the negative consequences of spanking" (p. 204), partly because trivial-to-small effects could be due to remaining trivial-to-small confounds that were unaccounted for. Some have claimed that even tiny effects can make a substantial difference in large populations (Rosnow & Rosenthal, 2003), but these claims were based on arguments now thought to be faulty. Furthermore, examples from medical science used to illustrate the "small is big" argument are now known to have been miscalculated or inapplicable (see Ferguson, 2009, for a through critique).

Presumably, it is Ferguson's use of *pr*s and Gershoff and Grogan-Kaylor's use of bivariate correlations that accounts for the discrepant effect sizes reported in these two most recent meta-analyses. The latter (citing Borenstein, Hedges, Higgins, & Rothstein, 2009) wrote: "Because meta-analyses are focused on simple effects . . . bivariate associations such as standardized differences of means or correlations were selected over adjusted coefficients from multivariate models" (p. 456). In contrast, consensus standards for rigorous meta-analyses (e.g., Reeves et al., 2008) prefer adjusted effects (e.g., β or *pr*), but acknowledge the difficulty of choosing among alternative adjusted effects. We are sympathetic to this "apples and oranges" problem and introduce two innovations herein to resolve it. Our innovations, thus, help fulfill the potential envisioned by Gershoff (2002b) when she said, "I sincerely hope that future meta-analyses of parental corporal punishment will have sufficient data on third variables to include them either as control or moderator variables" (p. 606).

The Present Study

We combined the strategies of the two most recent meta-analyses of spanking. Following Gershoff and Grogan-Kaylor (2016), we used weighted mean cross-sectional and longitudinal correlations between spanking and externalizing problems. We also collected a third correlation necessary to improve casual inferences of the effect of spanking on externalizing. For each included study, we obtained (a) a cross-sectional correlation between Time-1 (T1) spanking and T1 externalizing, (b) a longitudinal correlation between T1 spanking and T2 externalizing, and (c) a stability correlation between T1 externalizing and T2 externalizing. We expected that the weighted means of the first two correlations would approximate Gershoff and Grogan-Kaylor's effect sizes or be smaller due to our restriction to customary or open-handed spanking.

We used these three correlations to compute a mean standardized regression coefficient (β) for each study. This partial effect size is similar to Ferguson's (2013) partial *r*. When controlling for identical covariates, βs are typically smaller than partial *r*s by about |.01| in the range of values expected herein and have identical statistical tests (Pedhazur, 1997). We preferred βs over *pr*s because they are more familiar to readers and reported in more publications. We expected our average β to be slightly larger than its equivalent in Ferguson's (2013) meta-analysis, because his effect sizes also controlled for other covariates.

We go beyond Ferguson's (2013) methods by introducing two sensitivity tests to differentiate whether mean βs reflect residual confounding or actual causal effects. The first test we introduce, a slope prediction, computes the correlation between spanking and simple change scores in the outcome (e.g., the difference calculated by subtracting T1 externalizing from T2 externalizing, which is the *slope* predicted in a two-occasion linear growth model). This technique has not been used in prior meta-analyses of disciplinary spanking but is introduced here to check the robustness of effects obtained in the more typically employed cross-lagged "beta method."

Testing robustness with this "slope prediction method" is particularly helpful because it is usually biased in the opposite direction as the cross-lagged beta method. This has been shown, for example, in the pattern of results for many corrective actions for children, including out-of-home placements (Berger, Bruch, Johnson, James, & Rubin, 2009), four parental disciplinary responses, and two professional interventions (psychotherapy and Ritalin: Larzelere, Cox, & Smith, 2010; Larzelere, Ferrer, et al., 2010). Across both studies, *all* 13 significant effects predicting residualized change scores (the beta method) indicated that all corrective actions were harmful. In contrast, all 10 significant effects of simple change scores (the slope method) indicated beneficial effects.

The fact that analyses of these two types of change scores can produce contradictory results has been known at least since Lord's (1967) paradox. Although Cronbach and Furby (1970) favored analyses of residualized change scores over simple change scores, simple change scores have made a comeback (e.g., Allison, 1990; Rogosa & Willett, 1985) and are routinely used in linear growth models, Time × Treatment (2 × 2) repeated-measures analysis of variance, and differences in differences (Angrist & Pischke, 2009). Because statistical controls are imperfect, analyses of residualized change scores (betas) are generally biased *against* corrective actions (i.e., biased in the direction of the groups' initial differences in symptoms to be corrected). Because corrective actions are more likely to be used when problems get worse, analyses of simple change scores (slopes) are often biased *in favor of* corrective actions because they get credit for regression toward the mean (Larzelere, Ferrer, et al., 2010).

Because of these biases, Angrist and Pischke (2009) concluded that analyses of the two types of change scores might "have a useful bracketing property . . . bounding the causal effect of interest (given some assumptions about the nature of selection bias)" (pp. 245–246). But what those assumptions are, and how often they are met is not clear. For example, in a classic study in econometrics, LaLonde (1986) found that differences-in-differences analyses using the slope method were biased in favor of a corrective action (a job training program for the unemployed) *relative to* analysis of covariance-type analyses using the beta method, but *both* estimates were too pessimistic for men and both were too optimistic for women, compared with the "true" effect from the randomized part of the study. Nonetheless, Angrist and Pischke encouraged researchers to test robustness across both types of change scores as a promising way to distinguish stronger from weaker causal evidence, even if the resulting conclusiveness falls short of a randomized trial (see also Duncan, Engel, Claessens, & Dowsett, 2014). Even though the two types of change scores do not always bracket the true causal effect, robust consistency across the two change scores is a symptom of a true causal effect, whether accomplished by an ideal randomized study or by convincing propensity-score methods (Haviland, Nagin, & Rosenbaum, 2007).

The second sensitivity test we introduce, the backwards test, is a type of discriminant validity (or "falsification test" per Pizer, 2016) to check whether results replicate after reversing Times 1 and 2 (Galton, 1886). Residual confounding can work in reversed time as easily as forward in time, whereas actual causal effects can *only* operate forward in time. If a treatment was effective in a randomized trial, for example, running the analysis in reversed time would not replicate the effect, because the treatment and control groups would be similar on the pretest. Backwards tests can be confused with child effects, but differ from child effects in that child effects can occur only forward in time. Backwards tests and child effects both use the association between T1 externalizing and T2 spanking, but child-effect analyses control for T1 spanking, whereas backwards tests control for T2 externalizing.

Our final proposed technique, an "adjusted confound impact," tests whether potential confounds modify the central meta-analytic mean effect size. To do this, a unique 4 × 4 correlation matrix is constructed for each potential confound using correlations between the potential confound and the three variables in the original 3 × 3 correlation matrix already discussed. For example, eight of our included studies controlled for income. Hence, we computed a 4 × 4 correlation matrix for income, T1 spanking, T1 externalizing, and T2 externalizing. This 4 × 4 matrix was then used to estimate a meta-analytic standardized regression coefficient (β) that controlled for income as well as T1 externalizing.

In brief, then, this presentation consists of three sets of analyses: an approximation of the bivariate and partial effect sizes obtained by prior meta-analyses, two sensitivity tests to help discern whether the trivial-to-small effect sizes reported by Ferguson (2013) represent residual confounding or actual causal effects, and an assessment of several likely confounds to ascertain whether their inclusion impacts the mean effect sizes obtained from the first set of analyses.

Method

Several "Collaborations" have provided recent guidelines for improving the validity of causal inferences in meta-analyses of nonrandomized studies. These include the Cochrane Collaboration (Reeves et al., 2008), the Campbell Collaboration (Methods Group of the Campbell Collaboration, 2016), the American Psychological Association (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008), and related collaborative efforts (Wells et al., 2013). Where possible, we adhered to these guidelines.

Clarifying the Research Question

One guideline is explicitness in the research question. Our research question is simply *What are the effects of disciplinary spanking on subsequent externalizing behavior problems?* The Campbell Collaboration also requires further clarification of the precise intervention and "comparator," that is, the condition to which the intervention is being compared.

Our intervention is spanking, which has been defined as open-handed swats to the buttocks or extremities (Friedman & Schonberg, 1996). Because slapping is open handed and often delivered to the extremities, some slapping meets this definition, but we were concerned that slapping might also be delivered to the face, head, or torso, particularly for older children. Thus, we included studies that used a joint "spank or slap" question only for children under the age of 5. Otherwise we included studies that used the terms "spanking," "physical punishment," or "corporal punishment," assuming that the latter two phrases would generally be interpreted by parents as equivalent to spanking. We excluded studies that added other terms that made the operational definition too broad (e.g., push, shove, yell) or too severe (hit, use object).

Delineating the comparator proved more difficult. Randomized trials compare a well-defined treatment with a clear alternative, but what is the alternative in longitudinal studies of corrective actions such as spanking? Most longitudinal studies use a continuous measure of spanking frequency. Default linear statistics then primarily contrast the outcomes associated with the most versus the least spanking (i.e., overusage vs. nonusage). Another complication is that nonusage is often limited to the past week or month or an indefinite recent period of time. Few studies actually compare spanking with zero *lifetime* spanking. A meta-analysis can only work with the studies available, but it still needs to clarify the precise intervention and what it is being compared with, to prevent generalizing its results beyond that contrast.

Research Designs and Confounds

Another guideline relevant to causal inference is an explicit statement of the minimally acceptable research design. We based these meta-analyses on longitudinal designs that controlled at least for initial (i.e., T1) scores on a proxy of the outcome variable. The Collaboration guidelines also state that studies relevant to the research question with causal validity that approximates or exceeds the minimally acceptable designs should not be excluded (Reeves et al., 2008; Sterne et al., 2016). Because our primary focus is on enhancing causal inferences from longitudinal studies (and due to page limits), summaries of studies with causally stronger designs (randomized or quasi-experimental) or analyses that enhance causal inferences (propensity-score methods, within-person analyses) were relegated to the Supporting Information.

The Collaborations also require meta-analysts to identify the most important confounders in the planning stage (Sterne et al., 2016; Valentine & Thompson, 2013). The two types of covariates shown to be most useful for approximating unbiased causal estimates were pretest measures of the outcome variable and variables central to the treatment-selection process (Steiner, Cook, Shadish, & Clark, 2010). Accordingly, the most relevant confounds for our meta-analyses were (a) initial scores on the outcome variable and (b) persistent oppositional defiance, because it is likely to elicit a negative disciplinary consequence such as spanking.

We evaluated the likely importance of other potential confounds by using Frank's (2000) *confound impact*, the product of the two correlations of the potential confound z with the causal variable x and the outcome variable y (i.e., r_{xz} times r_{yz}). We estimated confound impact from three available longitudinal data sets previously used to study spanking. The mean correlations of each potential confound with T1 spanking and with T2 externalizing are listed in Table S1 in the Supporting Information. In these preliminary analyses, the largest confound impacts were for additional measures of externalizing problems, parenting stress, negative parenting, and positive parenting.

Literature Search and Coding

To locate relevant studies, we considered all studies in the most recent meta-analyses of spanking (Ferguson, 2013; Gershoff & Grogan-Kaylor, 2016). We searched the following databases for more recent publications from 2011 to early 2017: PsycInfo (225 references), PubMed (638), ERIC (128), and Sociological Abstracts (133). These terms were used: spank* or corporal punishment or physical punishment or harsh punishment or corporal disciplin* or physical disciplin* or harsh disciplin* (where * indicates a wildcard). Finally, we searched *Web of Science* for all recent articles (448) citing Gershoff (2013) or one of seven relevant literature reviews from 1996 to 2016 (including Horn, Joseph, & Cheng, 2004; Larzelere, 1996, 2000). All searches combined yielded 1,173 unique references from 2011 to early 2017.

Titles and abstracts were then screened to identity potentially relevant publications. Of the 284 potentially relevant articles identified, 247 were excluded for reasons shown in Table S2. Thirty-seven qualified for the meta-analysis. Because only three studies included data beyond two relevant occasions, we focused on the youngest two relevant occasions. Our "Time 1" (T1) was, thus, the youngest occasion with children older than 18 months that included measures of both spanking and externalizing or a reasonable proxy thereof, and "Time 2" (T2) was the next occasion with a measure of externalizing.

Following Lipsey and Wilson (2001), we then combined information from the same longitudinal cohort at the same ages when published in multiple studies. For example, 12 of the 37 studies analyzed data from the Fragile Families longitudinal data set. We counted publications from the same longitudinal data set as separate studies only if they focused on mostly different birth cohorts or different age groups. Otherwise we averaged available correlations across identical occasions or selected the publication with an available correlation matrix that was most relevant to our research question. We requested unpublished correlation matrices from several authors and are grateful to those who supplied them. This process yielded 14 distinct "sources" (i.e., individual articles or syntheses across overlapping publications), which are listed in Table 1.

All spanking was exclusively or mostly parent-reported. To minimize mono-source bias, we used externalizing measures from other sources when available. The authors collaborated in coding the studies' methods and other characteristics. Discrepancies were resolved by consensus.

Analyses

First, we computed random-effects weighted means (Borenstein et al., 2009, pp. 97–102) for the *three central correlations*: T1 spanking with T1 externalizing, T1 spanking with T2 externalizing, and T1 externalizing with T2 externalizing. We used these three correlations to calculate each study's standardized regression coefficient (β) for predicting T2 externalizing from T1 spanking, controlling for T1 externalizing. Then, we computed the overall weighted mean β. (The Supporting Information gives the equation and sample SPSS and Mplus syntax for calculating this β coefficient from the three correlations.)

We then conducted the two sensitivity tests. The first test, a slope prediction, calculated the effect of spanking on the "slope" of externalizing from T1 to T2 in a linear growth model. (This test used the same three correlations to predict the slope, equivalent to predicting the simple change score, using an exact equation or Mplus syntax, both shown in Supporting Information.) The second sensitivity test was the backwards test. This test repeated both the beta method and the slope prediction method, *after* reversing the two data occasions.

Finally, we calculated an adjusted confound impact for each potential confound, using ordinary least squares (OLS) regression analyses based on 4 × 4 correlation matrices to test how much the β predicting subsequent externalizing from spanking changed after adding each covariate. Whereas Frank's (2000) measure of confound impact provided a preliminary ranking of confounder importance from three commonly used data sets (Table S1), our adjusted confound impact went further by testing whether confounders influenced subsequent *change* in externalizing beyond their correlations with T1 externalizing in our 14 sources.

Table 1

Cross-Sectional, Longitudinal, and Stability Correlations for Spanking (x) *and Externalizing Problems* (y) *Used to Predict Residualized Gain Scores in Externalizing* (***β-Method***) *and Simple Gain Scores in Externalizing* (***Slope Method:*** $\boldsymbol{r_{x_1(y_2-y_1)}}$) *From Spanking*

Source	M_{age}[a]	N	$r_{x_1y_1}$	$r_{x_1y_2}$	$r_{y_1y_2}$	β	$r_{x_1(y_2-y_1)}$
Barnes, Boutwell, Beaver, and Gibson (2013)	4.1	550	.07$^{\text{ns}}$	.19	.48	**.16*****	**.12*****
Baumrind, Larzelere, and Owens (2010)	4.5	87	.35	.01$^{\text{ns}}$	.29	**−.10**	**−.29****
Berlin et al. (2009)	2.1	2573	.15	.12	.52	**.04***	**−.03**
Coley, Kull, and Carrano (2014)	3.4	581	.19	−.05$^{\text{ns}}$	.33	**−.12****	**−.21*****
Ellison et al. (2011)	3.0	456	.22	.14	.25	**.09***	**−.07**
Fragile Families[b]	3.0	3,575	.20	.18	.55	**.08*****	**−.02**
Gershoff, Lansford, Sexton, Davis-Kean, and Sameroff (2012)	6.2	11,044	.15	.15	.51	**.08*****	**.00**
Gershoff et al. (2016)	3.6	2,063	.25	.17	.50	**.05***	**−.08*****
Gunnoe and Mariner (1997)	7.8	1,112	.15	.15	.12	**.13*****	**.00**
Lansford, Wager, Bates, Dodge, and Pettit (2012)	5.0	585	.19	.21	.50	**.12****	**.02**
Larzelere, Ferrer, et al. (2010)	4.9	1,464	.28	.20	.51	**.07****	**−.07****
Mendez, Durtschi, Neppl, and Stith (2016)	2.5	218	.30	.21	.60	**.03**	**−.10**
Mulvaney and Mebert (2007)	3.0	979	.23	.22	.57	**.09*****	**−.01**
Straus, Sugarman, and Giles-Sims (1997)[c]	7.5	785	.27	.27	.49	**.15*****	**.00**
Sum, weighted mean *r*s & β[d]		26,072	.20	.16	.46	**.07*****[e]	**−.04***
I^2 (95% prediction intervals below)[f]			80.3	76.5	95.7	70.2	78.0

Note. The middle three data columns give the cross-sectional ($r_{x_1y_1}$) and longitudinal ($r_{x_1y_2}$) correlations between T1 spanking and T1 and T2 externalizing, respectively, followed by the stability correlations ($r_{y_1y_2}$) between T1 externalizing and T2 externalizing. The two right-hand columns give the standardized regression coefficients (β) predicting T2 externalizing from T1 spanking controlling for T1 externalizing, and the correlations ($r_{x_1(y_2-y_1)}$) between T1 spanking and the slope of externalizing from T1 to T2 (i.e., T2 externalizing minus T1 externalizing). The bold face in the two parenthetical-stated methods in the title correspond to the boldfaced columns.

[a] Mean age at the occasion treated as Time 1 (T1) for this meta-analysis. [b] Mean of correlations from Altschul, Lee, and Gershoff (2016), Gromoske and Maguire-Jack (2012), Lee, Altschul, and Gershoff (2013), and Maguire-Jack, Gromoske, and Berger (2012). [c] Using correlations from attempted duplication by Larzelere, Cox, et al. (2010). [d] The summation row presents random-effects mean *r*s and uses the same equations to estimate the random-effects mean βs (Borenstein et al., 2009). [e] Partial $r_{y_2x.y_1} = \frac{\sqrt{1-r^2_{xy_1}}}{\sqrt{1-r^2_{y_2y_1}}}\beta_{y_2x.y_1} = .08$. Semipartial $r_{y_2(x.y_1)} = \sqrt{1-r^2_{y_1x}}\beta_{y_2x.y_1} = .07$. [f] 95% prediction intervals (Borenstein et al., 2009) for estimated range of "true" effects in last five columns: (.09, .31), (.06, .26), (.22, .64), (−.02, .16), (−.15, .07).

$^*p < .05$. $^{**}p < .01$. $^{***}p < .001$ (in the two right-hand columns only). $^{ns}p > .05$ in middle three columns; otherwise, $p < .01$ for all *r*s in those columns.

Results

The cross-sectional ($r_{x_1y_1}$) and longitudinal ($r_{x_1y_2}$) correlations between spanking and externalizing and the outcome stability correlations ($r_{y_1y_2}$) for the 14 sources are listed in Table 1 along with the meta-analytically weighted means of those correlations. As shown in the third and fourth data columns, the mean cross-sectional ($r = .20$) and longitudinal correlations ($r = .16$) of Time-1 spanking with concurrent and later externalizing were similar to mean unadjusted correlations from previous meta-analyses ($d = .39$ [$= r = .19$], Gershoff, 2002a; $d = .21$ [$r = .10$], Paolucci & Violato, 2004; $d = .41$[$r = .20$], Gershoff & Grogan-Kaylor, 2016). This approximation to Gershoff and Grogan-Kaylor's (2016) correlations, despite limiting our studies specifically to customary spanking, suggests that unadjusted correlations do not vary much by physical punishment severity, consistent with one of their conclusions.

Standardized regression coefficients (β) predicting T2 externalizing from T1 spanking controlling for T1 externalizing are shown in the next-to-last column in Table 1. The mean effect size was β = .07 ($pr = .08$), $p < .001$, approximating Ferguson's (2013) effect size and his definition of trivial in size (i.e., smaller than Cohen's [1988] small effect size of $r = .10$).

The first sensitivity test was an attempt to show robustness by replicating the beta results with the slope method. The correlations between T1 spanking and the *slope* of externalizing problems from Time 1 to Time 2 appear in the last column of Table 1. The mean $r_{x_1(y_2-y_1)}$ was −.04, $p < .05$, indicating that spanking predicted significant *decreases* in externalizing from Time 1 to Time 2. As noted earlier, the slope prediction is a widely used alternative analysis of longitudinal change and is typically biased in the opposite direction as the beta method (e.g., Angrist &

Pischke, 2009). Viewed individually, one might be tempted to infer causation when any featured analysis meets conventional levels of significance. But when a robustness test produces results in *opposite* directions, a more reasonable inference may be that each statistic is impacted more by its own particular bias than by the actual causal effect under investigation. Put another way, a "causal relation" that is not robust enough to replicate across alternative methods of analysis may not represent an actual cause at all.

The second sensitivity test, the backwards test, was conducted by recalculating both types of analyses after reversing Times 1 and 2. That is, T2 spanking was used to predict T1 externalizing after controlling for T2 externalizing. Similarly, T2 spanking was used to predict reversed slopes from T2 back to T1 (i.e., T1 externalizing minus T2 externalizing). This was done for the eight studies with a measure of spanking at T2. In reversed time, T2 spanking predicted *higher* T1 externalizing scores, controlling for T2 externalizing, $\beta = .05$, $p < .001$, and it predicted *reversed* slopes in a linear growth model, $\beta = -.05$, $p < .01$. Thus, the backwards tests yielded the same contradictory results as analyses forward in time, suggesting that the significant associations between spanking and subsequent change in externalizing problems may be statistical artifacts due to residual confounding rather than actual causal effects. (Recall that residual confounding works in reverse; actual causal effects can *only* operate forward in time.)

The final set of analyses, the adjusted confound impacts, tested whether the potential confounding variables modified the apparent effect of spanking (Table 2). Adding an additional covariate barely changed the central cross-lagged path coefficient of $\beta = .074$ from spanking to subsequent externalizing, with two exceptions. Most of the adjusted coefficients ranged from $\beta = .067$ (controlling for scold or yell) to $\beta = .083$ (income). Only nonphysical punishment and the conceptually ambiguous variable "hostile-ineffective parenting" (aka "perceived child difficulty," because it fails to distinguish the child's tendency to exasperate from the exasperation the child elicited) resulted in larger changes (from $\beta = .074$ to $\beta = .096$ and to $\beta = .048$, respectively). For estimating the effect of spanking on subsequent externalizing problems, the addition of covariates did little beyond controlling for baseline externalizing. This finding differed from our expectations, but was consistent with Steiner et al.'s (2010) evidence that demographic variables do little to reduce self-selection bias.

Discussion

Methodological Issues

Due to the limited ability to employ randomized designs, most child development research is insufficient for conclusive causal inferences. "Starting from behind" as we are, it is imperative for developmental science to develop more rigorous techniques to differentiate between more versus less adequate *approximations* of valid causal inferences from longitudinal analyses. In keeping with this imperative, we have illustrated two sensitivity analyses, a slope prediction to test the robustness of a meta-analytic partial effect size (β) and a backwards test as a type of discriminant validity. Ferguson's (2013) study is the only previous meta-analysis of spanking that reported an effect size adjusted for initial scores on the outcome variable. We have improved on Ferguson's approach by averaging βs based on *identical* confounds (baseline externalizing plus other confounds one at a time), thus overcoming the "apples and oranges" problem inherent in averaging βs from discrepant sets of covariates.

Our first sensitivity test compared two alternative methods for analyzing change, often biased in opposite directions (against and in favor of corrective actions). Tests of robustness should be used more often in developmental science (Duncan et al., 2014), both in individual studies and meta-analyses. Our meta-analytic results indicated that the trivial partial effect sizes reported by Ferguson (2013) and replicated herein were more consistent with residual biases than with a true causal effect. Although simpler meta-analytic methods are easier to conduct and assimilate, they can produce tight confidence intervals around a systematic bias, producing a false sense of "spurious precision"—a significant, but incorrect effect size (Egger et al., 2001).

We also introduced an adjusted confound impact to assess whether potential cofounding variables would modify the overall partial effect size in a meta-analysis. This approach advances meta-analytic methods in two ways. First, it expands Frank's (2000) confound impact statistic by adjusting for initial differences in the outcome variable. If the confound impact of a variable is fully accounted for by its association with T1 outcome scores, then it is less important to control for in studies predicting change. Second, a meta-analytic estimate of the adjusted confound impact has more precision than an estimate based on a single study. Meta-analytic tests of potential confounders can inform future studies about which confounders to incorporate. Better understanding of confounders can lead to

Table 2

Modified Standardized Path Coefficients (β) From T1 Spanking to T2 Externalizing After Controlling for Potential Confounds One at a Time in Addition to T1 Externalizing

Potential confound	Number of sources that included the potential confound	β (T1 spanking predicting T2 externalizing)
(Unadjusted longitudinal correlation)	14	.164
(Cross-lagged β, controlling only for T1 externalizing)	14	.074
Cross-lagged β, controlling for T1 externalizing plus (one confound at a time) . . .		
Second measure of externalizing (hyperactivity or earlier emotionality)	7	.082
Parenting stress	2	.077
Maternal depression	3	.071
Maternal age at birth	1	.071
SES/income	8	.083
Male child	11	.078
Race (both entered simultaneously)	5	.066
African American	10	.073
Hispanic American	5	.076
Positive parenting		
Parental support (emotional support, cognitive stimulation, positive interaction)	6	.074
Disciplinary reasoning	2	.083
Nonphysical disciplinary consequences (privilege removal, sending child to room)	3	.096
Negative parenting		
Scold-yell	6	.067
Hostile-ineffective/perceived child difficulty[a]	1	.048

Note. T1 = Time 1; T2 = Time 2.

[a]Variable that combines anger/hostility with perceived child difficulty (from National Longitudinal Survey of Children and Youth).

cumulative improvement in causal inferences, which could then provide less biased causally relevant results in individual studies and meta-analyses.

Of course, meta-analyses are only as good as the evidence available from individual studies. Toward this end, longitudinal studies should provide the correlation matrices sufficient to replicate their results. It would be even better for studies to adopt methods to enhance the validity of causal inferences beyond the typical longitudinal study (Larzelere & Cox, 2013). Examples include randomized and quasi-experimental designs (Shadish, Cook, & Campbell, 2002), propensity-score methods (Haviland et al., 2007; Steiner et al., 2010), and analyzing natural experiments with instrumental variables (Angrist & Pischke, 2009).

The adoption of better methods can also improve research on other causally relevant research questions in developmental science by answering the following kinds of questions: What causal processes could produce the associations found in the research? Are the statistical assumptions for valid causal inferences satisfied? If not, can they be tested, or can sensitivity tests help determine whether the conclusions depend on those assumptions? What competing explanations can plausibly account for relevant research findings to date, and can they be tested against each other (Larzelere, Cox, & Swindle, 2015)? Child developmental research needs to deal with these questions in increasingly sophisticated ways, not only in meta-analyses, but in individual studies. Only then can the research provide the valid causal inferences necessary for optimal applications to improve human development.

Substantive Issues

Although our primary objective is to improve causal inferences in meta-analyses of longitudinal studies, our results have implications for the evaluation of disciplinary spanking as a corrective action. Like previous studies, we found that the best predictor of later externalizing was baseline externalizing. Once baseline externalizing was controlled for, demographics and most parenting characteristics had negligible impacts on the association between spanking and later externalizing (tested as covariates; moderators are tested in Supporting Information). These findings highlight the need to control for baseline maladjustment in longitudinal investigations of the effectiveness of any corrective disciplinary action.

In addition, we found that *neither* the significant "adverse" impact of spanking suggested by the usual β method or the significant "beneficial" impact of spanking suggested by slope predictions was robust enough to overcome the methodological bias of the alternative method. Put another way, it seems likely that the true average causal effect of spanking on externalizing is so close to zero that significant results require residual confounding (in the β method) or regression toward the mean (in slope predictions) in their favor. Although it is possible that the actual causal effect is either more detrimental or more beneficial than either of those estimates (see LaLonde, 1986), the lack of robustness is inconsistent with an unbiased causal estimate. Previous longitudinal analyses have found similar contrasting results for other corrective actions, including out-of-home placements (Berger et al., 2009) and alternatives that could replace spanking (Larzelere, Ferrer, et al., 2010; Larzelere, Cox, et al., 2010).

When average causal effects of corrective actions are too close to zero to document robustness, researchers should then try to differentiate effective versus counterproductive usages of such actions. The significant meta-regression results in Table S5 indicate that spanking becomes more harmful as child age increases, has become more effective in recent decades, and is more effective for American ethnic minority parents, ps < .05.

Because meta-analytic moderator tests are often under-powered for making these discriminations (see Table S6), recent guidelines also recommend using individual studies to suggest discriminations that can be missed by meta-analytic summary statistics (Valentine & Thompson, 2013). Although recent parenting research seems to focus more on broad categories (e.g., harsh parenting) than discriminations among specific tactics or clarifying how to use any tactic as effectively as possible, a few characteristics of spanking effectiveness have been identified. These include an intermediate level of usage (Lansford, Wager, Bates, Pettit, & Dodge, 2012; MacKenzie, Nicklas, Waldfogel, & Brooks-Gunn, 2013), phasing spanking out by age 9 or 11 (Ellison, Musick, & Holden, 2011; Gunnoe, 2013), and using spanking primarily as a back-up for milder disciplinary tactics, such as reasoning (Larzelere, Sather, Schneider, Larson, & Pike, 1998) or time-out (Roberts & Powers, 1990).

Unfortunately, this more nuanced knowledge can easily get lost in the usual meta-analytic emphasis on one overall effect size. Although well-conducted meta-analyses are useful tools for reducing error variance across studies asking roughly the same question—and the present study seeks to make them even *better* tools—results from meta-analyses must be considered in conjunction with results from innovative studies attempting to advance our understanding of corrective actions beyond an unconditional "thumbs up" or "thumbs down."

References

References marked with an asterisk contributed to the meta-analytic summaries in Tables 1 and 2. Other studies that met the inclusion criteria are listed in Table S3.

Allison, P. D. (1990). Change scores as dependent variables in regression analysis. In C. Clogg (Ed.), *Sociological methodology 1990* (pp. 93–114). Oxford, UK: Blackwell. https://doi.org/10.2307/271083

*Altschul, I., Lee, S. J., & Gershoff, E. T. (2016). Hugs, not hits: Warmth and spanking as predictors of child social competence. *Journal of Marriage and Family*, *78*, 695–714. https://doi.org/10.1111/jomf.12306

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's approach*. Princeton, NJ: Princeton University Press.

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology. *American Psychologist*, *63*, 839–851. https://doi.org/10.1037/0003-066X.63.9.839

*Barnes, J. C., Boutwell, B. B., Beaver, K. M., & Gibson, C. L. (2013). Analyzing the origins of childhood externalizing behavioral problems. *Developmental Psychology*, *49*, 2272–2284. https://doi.org/10.1037/a0032061

*Baumrind, D., Larzelere, R. E., & Owens, E. B. (2010). Effects of preschool parents' power assertive patterns and practices on adolescent development. *Parenting: Science and Practice*, *10*, 157–201. https://doi.org/10.1080/15295190903290790

Berger, L. M., Bruch, S. K., Johnson, E. I., James, S., & Rubin, D. (2009). Estimating the "impact" of out-of-home placement on child well-being: Approaching the problem of selection bias. *Child Development*, *80*, 1856–1876. https://doi.org/10.1111/j.1467-8624.2009.01372.x

*Berlin, L. J., Ispa, J. M., Fine, M. A., Malone, P. S., Brooks-Gunn, J., Brady-Smith, C., & Bai, Y. (2009). Correlates and consequences of spanking and verbal punishment for low-income White, African American, and Mexican American toddlers. *Child Development*, *80*, 1403–1420. https://doi.org/10.1111/j.1467-8624.2009.01341.x

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. West Sussex, UK: Wiley. https://doi.org/10.1002/9780470743386

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

*Coley, R. L., Kull, M. A., & Carrano, J. (2014). Parental endorsement of spanking and children's internalizing and externalizing problems in African American and Hispanic families. *Journal of Family Psychology*, *28*, 22–31. https://doi.org/10.1037/a0035272

Cronbach, L. J., & Furby, L. (1970). How we should measure change–or should we? *Psychological Bulletin*, *74*, 32–49. https://doi.org/10.1037/h0029382

Deptula, D. P., Henry, D. B., & Schoeny, M. E. (2010). How can parents make a difference? Longitudinal associations with adolescent sexual behavior. *Journal of Family Psychology*, *24*, 731–739. https://doi.org/10.1037/a0021760

Duncan, G. J., Engel, M., Claessens, A., & Dowsett, C. J. (2014). Replication and robustness in developmental research. *Developmental Psychology*, *50*, 2417–2425. https://doi.org/10.1037/a0037996

Egger, M., Smith, G. D., & Schneider, M. (2001). Systematic reviews of observational studies. In M. Egger, G. D. Smith & D. G. Altman (Eds.), *Systematic reviews in health care: Meta-analysis in context* (pp. 211–227). Cornwall, UK: BMJ Books. https://doi.org/10.1002/9780470693926.ch12

*Ellison, C. G., Musick, M. A., & Holden, G. W. (2011). Does conservative Protestantism moderate the association between corporal punishment and child outcomes? *Journal of Marriage and Family*, *73*, 946–961. https://doi.org/10.1111/j.1741-3737.2011.00854.x

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, *40*, 532–538. https://doi.org/10.1037/a0015808

Ferguson, C. J. (2013). Spanking, corporal punishment and negative long-term outcomes: A meta-analytic review of longitudinal studies. *Clinical Psychology Review*, *33*, 196–208. https://doi.org/10.1016/j.cpr.2012.11.002

Frank, K. A. (2000). Impact of a confounding variable on a regression coefficient. *Sociological Methods & Research*, *29*, 147–194. https://doi.org/10.1177/0049124100029002001

Friedman, S. B., & Schonberg, S. K. (1996). Consensus statements [from the invitational conference, the short- and long-term consequences of corporal punishment]. *Pediatrics*, *98*, 852–853.

Galton, F. (1886). Regression toward mediocrity in hereditary stature. *Journal of the Anthropological Institute of Great Britain and Ireland*, *15*, 246–263. https://doi.org/10.2307/2841583

Gershoff, E. T. (2002a). Corporal punishment by parents and associated child behaviors and experiences: A meta-analytic and theoretical review. *Psychological Bulletin*, *128*, 539–579. https://doi.org/10.1037/0033-2909.128.4.539

Gershoff, E. T. (2002b). Corporal punishment, physical abuse, and the burden of proof: Reply to Baumrind, Larzelere, and Cowan (2002), Holden (2002), and Parke (2002). *Psychological Bulletin*, *128*, 602–611. https://doi.org/10.1037/0033-2909.128.4.602

Gershoff, E. T. (2013). Spanking and child development: We know enough now to stop hitting our children. *Child Development Perspectives*, *7*, 133–137. https://doi.org/10.1111/cdep.12038

*Gershoff, E. T., Ansari, A., Purtell, K. M., & Sexton, H. R. (2016). Changes in parents' spanking and reading as mechanisms for Head Start impacts on children. *Journal of Family Psychology*, *30*, 480–491. https://doi.org/10.1037/fam0000172

Gershoff, E. T., & Grogan-Kaylor, A. (2016). Spanking and child outcomes: Old controversies and new meta-analyses. *Journal of Family Psychology*, *30*, 453–469. https://doi.org/10.1037/fam0000191

*Gershoff, E. T., Lansford, J. E., Sexton, H. R., Davis-Kean, P., & Sameroff, A. J. (2012). Longitudinal links between spanking and children's externalizing behaviors in a national sample of White, Black, Hispanic, and Asian American families. *Child Development*, *83*, 838–843. https://doi.org/10.1111/j.1467-8624.2011.01732.x

Giles-Sims, J., Straus, M. A., & Sugarman, D. B. (1995). Child, maternal, and family characteristics associated with spanking. *Family Relations*, *44*, 170–176. https://doi.org/10.2307/584804

*Gromoske, A. N., & Maguire-Jack, K. (2012). Transactional and cascading relations between early spanking and children's social-emotional development. *Journal of Marriage & Family*, *74*, 1054–1068. https://doi.org/10.1111/j.1741-3737.2012.01013.x

Gunnoe, M. L. (2013). Associations between parenting style, physical discipline, and adjustment in adolescents' reports. *Psychological Reports: Disability & Trauma*, *112*, 933–975. https://doi.org/10.2466/15.10.49.PR0.112.3.933-975

*Gunnoe, M. L., & Mariner, C. L. (1997). Toward a developmental-contextual model of the effects of parental spanking on children's aggression. *Archives of Pediatrics and Adolescent Medicine*, *151*, 768–775. https://doi.org/10.1001/archpedi.1997.02170450018003

Haviland, A. M., Nagin, D. S., & Rosenbaum, P. R. (2007). Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychological Methods*, *12*, 247–267. https://doi.org/10.1037/1082-989X.12.3.247

Herbst, C. M., & Tekin, E. (2014). Child care subsidies, maternal health, and child-parent interactions: Evidence from three nationally representative datasets. *Health Economics*, *23*, 894–916. https://doi.org/10.1002/hec.2964

Hill, N. E., & Tyson, D. F. (2009). Parental involvement in middle school: A meta-analytic assessment of the strategies that promote achievement. *Developmental Psychology*, *45*, 740–763. https://doi.org/10.1037/a0015362

Holden, G. W., Ashraf, R., Brannan, E., & Baker, P. (2016). The emergence of "positive parenting" as a revived paradigm: Theory, processes, and evidence. In D. Narvaez, J. M. Braungart-Rieke, L. E. Miller-Graff, L. T. Gettler, & P. D. Hastings (Eds.), *Contexts for young child flourishing* (pp. 201–214). New York, NY: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190237790.003.0010

Horn, I. B., Joseph, J. G., & Cheng, T. L. (2004). Nonabusive physical punishment and child behavior among African-American children: A systematic review. *Journal of the National Medical Association*, *96*, 1162–1168.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, *76*, 604–620.

*Lansford, J. E., Wager, L. B., Bates, J. E., Dodge, K. A., & Pettit, G. S. (2012). Parental reasoning, denying privileges, yelling, and spanking: Ethnic differences and associations with child externalizing behavior. *Parenting: Science and Practice*, *12*, 42–56. https://doi.org/10.1080/15295192.2011.613727

Lansford, J. E., Wager, L. B., Bates, J. E., Pettit, G. S., & Dodge, K. A. (2012). Forms of spanking and children's externalizing behaviors. *Family Relations*, *61*, 224–236. https://doi.org/10.1111/j.1741-3729.2011.00700.x

Larzelere, R. E. (1996). A review of the outcomes of parental use of nonabusive or customary physical punishment. *Pediatrics*, *98*, 824–828.

Larzelere, R. E. (2000). Child outcomes of nonabusive and customary physical punishment by parents: An updated literature review. *Clinical Child and Family Psychology Review*, *3*, 199–221. https://doi.org/10.1023/A:1026473020315

Larzelere, R. E., & Cox, R. B., Jr. (2013). Making valid causal inferences about corrective actions by parents from longitudinal data. *Journal of Family Theory & Review*, *5*, 282–299. https://doi.org/10.1111/jftr.12020

*Larzelere, R. E., Cox, R. B., Jr., & Smith, G. L. (2010). Do nonphysical punishments reduce antisocial behavior more than spanking? A comparison using the strongest previous causal evidence against spanking. *BMC Pediatrics*, *10*. https://doi.org/10.1186/1471-2431-10-10

Larzelere, R. E., Cox, R. B., Jr., & Swindle, T. M. (2015). Many replications do not causal inferences make: The need for critical replications to test competing explanations of non-randomized studies. *Perspectives on Psychological Science*, *10*, 380–389. https://doi.org/10.1177/1745691614567904

*Larzelere, R. E., Ferrer, E., Kuhn, B. R., & Danelia, K. (2010). Differences in causal estimates from longitudinal analyses of residualized versus simple gain scores: Contrasting controls for selection and regression artifacts. *International Journal of Behavioral Development*, *34*, 180–189. https://doi.org/10.1177/0165025409351386

Larzelere, R. E., & Kuhn, B. R. (2005). Comparing child outcomes of physical punishment and alternative disciplinary tactics: A meta-analysis. *Clinical Child and Family Psychology Review*, *8*, 1–37. https://doi.org/10.1007/s10567-005-2340-z

Larzelere, R. E., Sather, P. R., Schneider, W. N., Larson, D. B., & Pike, P. L. (1998). Punishment enhances reasoning's effectiveness as a disciplinary response to toddlers. *Journal of Marriage and the Family*, *60*, 388–403. https://doi.org/10.2307/353856

*Lee, S. J., Altschul, I., & Gershoff, E. T. (2013). Does warmth moderate longitudinal associations between maternal spanking and child aggression in early childhood? *Developmental Psychology*, *49*, 2017–2028. https://doi.org/10.1037/a0031630

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin*, *68*, 304–305. https://doi.org/10.1037/h0025105

MacKenzie, M. J., Nicklas, E., Waldfogel, J., & Brooks-Gunn, J. (2013). Spanking and child development across the first decade of life. *Pediatrics*, *132*, e1118–e1125. https://doi.org/10.1542/peds.2013-1227

*Maguire-Jack, K., Gromoske, A. N., & Berger, L. M. (2012). Spanking and child development during the first 5 years of life. *Child Development*, *83*, 1960–1977. https://doi.org/10.1111/j.1467-8624.2012.01820.x

*Mendez, M., Durtschi, J., Neppl, T. K., & Stith, S. M. (2016). Corporal punishment and externalizing behaviors in toddlers: The moderating role of positive and harsh parenting. *Journal of Family Psychology*, *30*, 887–895. https://doi.org/10.1037/fam0000187

Methods Group of the Campbell Collaboration. (2016). *Methodological expectations of Campbell Collaboration intervention reviews: Conduct standards*. Oslo, Norway: Author.

*Mulvaney, M. K., & Mebert, C. J. (2007). Parental corporal punishment predicts behavior problems in early childhood. *Journal of Family Psychology*, *21*, 389–397. https://doi.org/10.1037/0893-3200.21.3.389

Paolucci, E. O., & Violato, C. (2004). A meta-analysis of the published research on the affective, cognitive, and behavioral effects of corporal punishment. *Journal of Psychology*, *138*, 197–221. https://doi.org/10.3200/JRLP.138.3.197-222

Pedhazur, E. J. (1997). *Multiple regression in behavioral research*. Toronto, CA: Wadsworth.

Pizer, S. D. (2016). Falsification testing of instrumental variables methods for comparative effectiveness research. *Health Services Research*, *51*, 790–811. https://doi.org/10.1111/1475-6773.12355

Reeves, B. C., Deeks, J. J., Higgins, J. P. T., & Wells, G. A. (2008). Including non-randomized studies. In J. P. T. Higgins & S. Green (Eds.), *Cochrane handbook for systematic reviews of interventions* (pp. 391–432). Chichester, UK: Wiley. https://doi.org/10.1002/9780470712184.ch13

Roberts, M. W., & Powers, S. W. (1990). Adjusting chair timeout enforcement procedures for oppositional children. *Behavior Therapy*, *21*, 257–271. https://doi.org/10.1016/S0005-7894(05)80329-6

Rogosa, D. R., & Willett, J. B. (1985). Understanding correlates of change by modeling individual differences in growth. *Psychometrika*, *50*, 203–228. https://doi.org/10.1007/BF02294247

Rosnow, R. L., & Rosenthal, R. (2003). Effect sizes for experimenting psychologists. *Canadian Journal of Experimental Psychology/Revue Canadienne De Psychologie Expérimentale*, *57*, 221–237. https://doi.org/10.1037/h0087427

Rothman, K. J., Greenland, S., & Lash, T. L. (2008). *Modern epidemiology* (3rd ed.). Philadelphia, PA: Wolter Kluwer.

Ryan, R. M., Kalil, A., Ziol-Guest, K. M., & Padilla, C. (2016). Socioeconomic gaps in parents' discipline strategies from 1988 to 2011. *Pediatrics*, *138*. https://doi.org/10.1542/peds.2016-0720

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

Society for Research in Human Development. (2013). *Resolution on corporal punishment (CP) of children*. Retrieved from http://www.srhdonline.org/resolution-on-corporal-punishment.html

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods*, *15*, 250–267. https://doi.org/10.1037/a0018719

Sterne, J. A. C., Hernan, M. A., Reeves, B. C., Savovic, J., Berkman, N. D., Viswanathan, M., . . . Higgins, J. P. T. (2016). ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. *British Medical Journal*, *355*(i4919), 1–7. https://doi.org/10.1136/bmj.i4919

Straus, M. A., & Stewart, J. H. (1999). Corporal punishment by American parents: National data on prevalence, chronicity, severity, and duration, in relation to child and family characteristics. *Clinical Child and Family Psychology Review*, *2*, 55–70. https://doi.org/10.1023/A:1021891529770

*Straus, M. A., Sugarman, D. B., & Giles-Sims, J. (1997). Spanking by parents and subsequent antisocial behavior of children. *Archives of Pediatrics and Adolescent Medicine*, *151*, 761–767. https://doi.org/10.1001/archpedi.1997.02170450011002

Valentine, J. C., & Thompson, S. G. (2013). Issues relating to confounding and meta-analysis when including nonrandomized studies in systematic reviews on the effects of interventions. *Research Synthesis Methods*, *4*, 26–35. https://doi.org/10.1002/jrsm.1064

Webster-Stratton, C., Reid, M. J., & Beauchaine, T. P. (2013). One-year follow-up of combined parent and child intervention for young children with ADHD. *Journal of Clinical Child and Adolescent Psychology*, *42*, 251–261. https://doi.org/10.1080/15374416.2012.723263

Wells, G. A., Shea, B., Higgins, J. P. T., Sterne, J., Tugwell, P., & Reeves, B. C. (2013). Checklists of methodological issues for review authors to consider when including non-randomized studies in systematic reviews. *Research Synthesis Methods*, *4*, 63–77. https://doi.org/10.1002/jrsm.1077

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's website:

Figure S1. Forest Plot for Beta Method, Predicting Residualized Change in Externalizing Behavior Problems From T1 Spanking

Figure S2. Forest Plot for Slope Prediction Method, Predicting Simple Change in Externalizing Behavior Problems From T1 Spanking

Figure S3. Scatterplots of Significant Continuous Moderators From Meta-Regression Results

Table S1. Preliminary Estimates of Confound Impacts ($r \times r$) of Potential Confounds

Table S2. Reasons That Potentially Relevant Articles Were Excluded Based on Their Full Text

Table S3. Reasons Why Other Publications Meeting the Inclusion Criteria Did Not Contribute to the Featured Meta-analysis

Table S4. Additional Characteristics of Studies Included in the Featured Meta-analysis

Table S5. Meta-Regression Results (Moderation Tests of Continuous Moderators)

Table S6. Moderation Tests of Categorical Moderators

Table S7. Methodological Criteria Relevant to Causal Validity From the Campbell Collaboration and the American Psychological Association

Appendix S1. Statistical Syntax Examples

Appendix S2. Derivation of Equation to Calculate the Standardized β for Predicting the Simple Gain Score ($y_2 - y_1$) from x Based on the Three Correlations Between x, y_1, and y_2

Appendix S3. Steps to Implement a Meta-analysis of Longitudinal Studies

Data S1. Calculating Fixed- and Random-Effects Weighted Means of rs for Longitudinal Meta-analyses