DOI: 10.1111/cdev.13701

EMPIRICAL REPORT

CHILD DEVELOPMENT

Spanking and externalizing problems: Examining within-subject associations

Joshua Pritsker

San Diego Mesa College, San Diego, California, USA

Correspondence
Joshua Pritsker, Purdue University, Psychological Sciences, 703 Third Street, West Lafayette, IN 47907, USA.
Email: jpritsk@purdue.edu

Abstract

This study examined the effects of spanking on externalizing on a within-subject level, while excluding causally irrelevant between-subject variance. Results from two longitudinal studies which used participants from the Child Development Project (n = 585) were reanalyzed with a random-intercept cross-lagged panel model using yearly measurements over ages 6–8. After removing between-subject variance, there were no significant effects of general spanking on externalizing (β = .06, .07). However, when done without objects and at a rate of about once per month or less, spanking showed beneficial effects (β = −.17, −.21). Results suggest that previous findings may be due to a failure to separate between-subject and within-subject variance. Additionally, results illustrate the need to examine limited spanking separately from more general forms of physical punishment.

INTRODUCTION

Spanking is a common parenting practice; within the 2018 results from the General Social Surveys, roughly two-thirds of Americans agreed that "it is sometimes necessary to discipline a child with a good, hard spanking" (Smith et al., 2018). While it's generally used with the intent to improve behavior, its true effects have been a topic of considerable debate (cf. Benjet & Kazdin, 2003). Of particular focus has been its effects on externalizing problems, such as aggression or defiance. Following research findings regarding correlations between physical punishment and externalizing (e.g., Gershoff & Grogan-Kaylor, 2016), many have warned that spanking leads to increases in externalizing problems. In contrast, some have argued that such results are misleading because of methodological problems (e.g., Larzelere et al., 2019). Most of the results come from methods that fail to properly separate between-subject and within-subject differences (elucidated below), and few studies have examined limited spanking exclusively from harsher forms of physical punishment.

Abbreviations: ALT-SR, autoregressive latent trajectory model with structured residuals; CDP, Child Development Project; CFI, comparative fit index; RI-CLPM, random-intercept cross-lagged panel model; RMSEA, root mean square error of approximation; SRMR, standardized root mean square residual.

Distinguishing sources of variance

Longitudinal observations contain two levels of variance: variance between subjects and variance within subjects. In this case, between-subject variance represents how parent-child dyads differ across each other, while within-subject variance represents differences within each dyad over time. Methodologists have noted that the level of focus for causal inference is typically the within-subject level rather than the between-subject one (Berry & Willoughby, 2016; Hoffman & Stawski, 2009). Despite this, commonly used longitudinal methods, such as cross-lagged panel models (CLPMs), fail to properly separate between-subject differences from within-subject ones (Berry & Willoughby, 2016; Curran et al., 2014; Hamaker et al., 2015). Instead, CLPMs mix these together, preventing one from examining the relations of interest. Using these mixed results can lead to misleading inferences by confounding the within-subject relations of interest with between-subject ones (i.e., Simpson's

Paradox), a problem that has been illustrated both analytically and through simulation (Berry & Willoughby, 2016; Hamaker et al., 2015; Usami, Todo, et al., 2019). For instance, it may be that children who are spanked more frequently have higher levels of externalizing problems between parent-child dyads. However, this does not imply that the same holds true within parent-child dyads over time. A synthetic example of this problem is given in Figure 1, where the effect sign differs depending on the level of inference.

A potential solution to this is to use difference scores, which implicitly remove preexisting between-subject differences. Following this, Larzelere et al. (2018) conducted a meta-analysis using both a two-wave difference model and cross-lagged correlations. They found adverse effects of spanking on externalizing when using typical cross-lagged correlations, but found beneficial effects when using difference scores. However, while this difference model does exclude between-subject differences, it assumes no instability and thus results may be confounded with regression toward the mean (Larzelere et al., 2010, 2018). Alternatively, one could directly model the two levels of variance separately by including subject-specific intercepts, explicitly modeling systematic differences while allowing for instability to be modeled. Building upon Curran et al. (2014), Berry and Willoughby (2016) did this, also adding latent slopes, resulting in what they referred to as the autoregressive latent trajectory model with structured residuals (ALT-SR). Using this model, they found no significant effects of physical punishment on subsequent externalizing problems in two separate datasets.

Separating forms of physical punishment

A second problem which has frequently been mentioned is that research has generally failed to adequately separate harsh and mild forms of physical punishment (Baumrind et al., 2002; Ferguson, 2013; Larzelere & Baumrind, 2010). As a result of this, assuming that harsh forms have adverse effects, milder forms may still have null or even beneficial effects while the grouped results suggest adverse effects. Even when limited forms of physical punishment are examined specifically, they are often examined independently instead of exclusively. For instance, although Larzelere et al. (2018) intended to focus on open-handed spanking, and excluded studies that explicitly included the use of objects, they noted that only one of their included studies *exclusively* focused on open-handed spanking (viz., Lansford et al., 2012). Examining mild forms of physical punishment independently rather than exclusively mixes them with harsher forms, again preventing proper estimation of effect sizes. Furthermore, measures of spanking often ask only about usage in the week before interview. This skews responses toward more frequent spanking, as only those who spank frequently are likely to have done so in a given week (Larzelere & Baumrind, 2010). As a result, studies unintentionally ignore infrequent spanking.

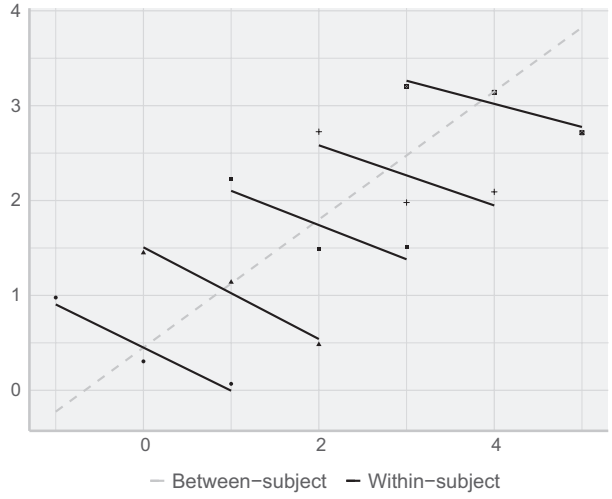

FIGURE 1 Example of opposing between-subject and within-subject associations

The present study

The present study reanalyzes the results of Lansford et al. (2012) and Study 1 of Lansford et al. (2011) while separating between-subject differences from within-subject ones and examining spanking done without objects at a frequency of less than once per week separately from general spanking. This reanalysis is important given the combined need to disaggregate the two levels of variance and rarity of research focusing on explicitly limited forms of spanking. Between-subject variance was separated from within-subject variance through the use of random-intercepts (RIs) with structured residuals in a RI-CLPM (Hamaker et al., 2015). Like the ALT-SR, but advancing beyond simple fixed effects and two-wave difference models, this model incorporates the multilevel nature of longitudinal data while allowing for estimation of cross-lagged paths and control for (in)stability.

There were two primary research questions being addressed. First, this study examines the effects of spanking frequency on externalizing behaviors in general while properly removing between-subject variance from effect estimates. Second, this study examines the effects of limited spanking on externalizing problems after excluding spanking that is done with objects or at a high frequency. Berry and Willoughby (2016) found no significant effects of general spanking and physical punishment on subsequent externalizing behaviors under a similar model, and I expected the same here. Limited spanking, however, has not been examined with a similar model. The original results of the studies reanalyzed here were discordant from each other, with the effect of limited spanking being nonsignificant and the effects of general spanking

being adverse. Recent meta-analyses have noted a lack of differences in bivariate correlations across severity levels (Gershoff & Grogan-Kaylor, 2016; Larzelere et al., 2018), finding adverse effects despite the level of severity. Larzelere et al. (2018) also failed to find significant differences in cross-lagged correlations across studies that used different definitions of a lack of spanking (e.g., no spanking in past week, past month, etc.), but did find differences when comparing the highest spanking score. These effect estimates, however, were still adverse except when characterized by endorsement rather than usage. Again, it should be emphasized that these previous results do not necessarily apply on a within-subject level. Given the lack of prior focus on within-subject effects and limited spanking, and no strong directional hypotheses, the present study could be viewed as conceptually more exploratory. Despite this, there were specific parameters of focus, and the methods used are appropriate for a confirmatory setting.

METHOD

Data sources

Descriptive data from Lansford et al. (2012) and from Study 1 of Lansford et al. (2011) were reanalyzed. These studies took their participants from the Child Development Project (CDP; Dodge et al., 1990), a multisite longitudinal study using participants recruited from Nashville and Knoxville, Tennessee and Bloomington, Indiana. Parents were approached at random during kindergarten preregistration for recruitment. Lansford et al. (2011) used all participants with relevant data, while Lansford et al. (2012) focused on a subgroup of participants who exclusively spanked in what was deemed to be a "mild" manner. As both used the same pool of participants, the mild sample from Lansford et al. (2012) may be viewed as a subset of the general sample from Lansford et al. (2011). These studies were chosen because of several advantages, including availability, having a sufficient number of measurement periods for model estimation, having evenly spaced measurement periods, allowing for the examination of the effects of limited spanking separately from overall physical punishment, and focusing on middle childhood rather than adolescence (Benjet & Kazdin, 2003; Larzelere et al., 2018). Within Lansford et al. (2011), detrimental effects were found independently for both spanking done with one's hand and with objects, while Lansford et al. (2012) found no significant cross-lagged effects once spanking was exclusively limited to those who only spanked without objects and at a frequency of once per month or less. The CDP sample included 585 children, of which 52% were categorized as male and 48% as female. Eight-one percent of these children were cataloged as being European American, 17% as African American, and 2% as being of other ethnic descent. 26% of the included families were headed by single mothers. Lansford et al. (2011) reported that during the first assessment at age 5, the mean socioeconomic status using the four-factor index from Hollingshead (1975) was 39.53 (SD = 14.01), corresponding to a status rating of "skilled craftsmen, clerical, and sales workers." After excluding those who had no data on spanking or externalizing, 562 children remained and were included in Lansford et al. (2011). Only those who spanked infrequently and without objects were included in Lansford et al. (2012), resulting in 258 participants.

Measures

Measures were taken from parents and teachers of the children on a yearly basis while the children were aged 6, 7, and 8. Lansford et al. (2011) also examined age 9 spanking and externalizing. However, age 9 measurements were not examined here as they were not included in Lansford et al. (2012). Furthermore, only 40% of the sample had data for age 9, although this was pre-planned rather than being due to attrition. Correlations, means, and SDs of the measures may be found in Supporting Information (Table S1).

Spanking

"Spanking" was used to refer to physical punishment by striking a child's buttocks with one's hand or an object. Spanking was measured by the frequency with which mothers reported that they spanked their child in the past year with their hand or with an object on a scale from 0 to 4, with *never* coded as 0, *less than once a month* as 1, *about once a month* as 2, *about once a week* as 3, and *about every day* as 4. In Lansford et al. (2012) those who used objects were excluded, but in Lansford et al. (2011) both items were averaged.

Externalizing problems

Externalizing problems were measured using items from the Teacher's Report Form of the Achenbach System of Empirically Based Assessment (Achenbach & Edelbrock, 1986). Internal consistency of this measure was reported to be high, with αs over .90 for all times. Lansford et al. (2011) reported the specific values of α = .95 for age 6 and α = .96 for both ages 7 and 8. Although the raw data were identical, comparisons across the studies should be done with standardized rather than unstandardized measures due to differences in data handling. Lansford et al. (2011) reported that means were comparable to previously reported values from a normative sample of non-referred children of the same ages. They also reported that they square-root transformed this measure.

TABLE 1 Results from random-intercept cross-lagged panel model

	Limited: Lansford et al. (2011), *n* = 258				General: Lansford et al. (2012), *n* = 563			
Path	***B***	***SE***	***p***	***β***	***B***	***SE***	***p***	***β***
$spank_{t-1} \rightarrow spank_t$[a]	.17	0.10	.079	.16 .17	.07	0.08	.365	.08 .07
$extern_{t-1} \rightarrow extern_t$[a]	−.01	0.10	.922	−.01 −.01	.11	0.07	.114	.10 .13
$spank_{t-1} \rightarrow extern_t$[a]	−2.61	1.10	.018	−.17 −.21	.15	0.12	.224	.07 .06
$extern_{t-1} \rightarrow spank_t$[a]	−.00	0.01	.366	−.07 −.06	.01	0.02	.604	.03 .03
$spank_t \leftrightarrow extern_t$[b]	−.21 −.41 −.41	0.20 0.15 0.15	.281 .281 .008	−.10 −.19 −.22	−.01 −.01 −.01	0.04 0.03 0.03	.800 .819 .819	−.02 −.01 −.01
$spank_{intercept} \leftrightarrow extern_{intercept}$	.59	0.21	.004	.33	.28	0.05	.000	.34

[a] For *β*: ages 6 → 7,7 → 8.
[b] Ages 6, 7, 8.

Harshness

Analyses were done upon both the complete sample using data from Lansford et al. (2011) and the limited subsample using data from Lansford et al. (2012). As reported in Lansford et al. (2012), the limited group excluded those who were spanked with objects or at a frequency of about once per week or more. Participants who failed to meet this criteria at any of the measurements at ages 6, 7, or 8 were excluded from the limited group, resulting in a total of 258 remaining participants. The remaining spanking could be viewed as somewhat similar to "customary spanking" as used by Larzelere et al. (2018), but with explicit restrictions rather than an assumption that people will respond to questions about "spanking" with such restrictions in mind. The general group from Lansford et al. (2011) had no such restrictions.

Analytic plan

Statistical analyses were done using the R language with the *lavaan* package (R Core Team, 2019; Rosseel, 2012). I derived covariance matrices using the reported correlation matrices and *SD*s, then fit an RI-CLPM (Hamaker et al., 2015) to these matrices and the reported means to examine within-subject effects. The RI-CLPM may be viewed as an extension of the standard autoregressive CLPM, adding random intercepts in order to separate between-subject differences from within-subject ones. The ALT-SR used by Berry and Willoughby (2016) is similar to this model, but includes latent slopes instead of time-specific means and requires a fourth wave unless additional assumptions are introduced. Models such as the RI-CLPM have been shown to remove the bias present in CLPMs (Berry & Willoughby, 2016; Hamaker et al., 2015), and are thus likely to better approximate causal effects.

I started with a model containing cross-sectional covariances, stability paths, and cross-lagged effects. As is typical with cross-lagged models, cross-sectional relations were specified as residual covariances rather than monodirectional paths. Following Hamaker et al. (2015), I included time-specific grand means. It should be noted that the interpretation of the stability paths here differs from in typical CLPMs, as here they represent the stability of the within-person deviations from one's baseline over time, whereas in typical CLPMs they represent the rank-order stability of individuals over time.

Cross-lagged and stability paths were constrained to be equal over time, as well as residual covariances. I tested to see if removing any of these constraints lead to significant improvement in model fit, and removed any that did. The final model, derived using this procedure on the covariance matrix from the larger sample of Lansford et al. (2011), was then fitted to the matrix from Lansford et al. (2012). For comparison, the same model was specified as a typical CLPM by constraining the variances and covariances of the random intercepts to 0.

RESULTS

None of the tested modifications to the model constraints resulted in a significant change in fit at $p < .05$, and thus none were kept. The model provided good fit when using the results from the general sample ($\chi^2(6, n = 563) = 8.81$, $p = .184$, standardized root mean square residual [SRMR] = .02, root mean square error of approximation [RMSEA] = .03, comparative fit index [CFI] = 1.00), as well as when using the subsample restricted to limited usage ($\chi^2(6, n = 258) = 8.41$, $p = .210$, SRMR = .03, RMSEA = .04, CFI = .99). In contrast, specifying the model as a typical CLPM yielded comparatively poor fit (General: $\chi^2(9, n = 563) = 133.00$, $p_{\chi^2} < .001$, SRMR = .06, RMSEA = .16, CFI = .90, $p_{\Delta\chi^2} < .001$; Limited: χ^2(9,

n = 258) = 65.41, $p_{\chi^2} < .001$, SRMR = .06, RMSEA = .16, CFI = .86, $p_{\Delta\chi^2} < .001$).

Results from the RI-CLPMs are detailed in Table 1. Substantial between-subject associations were present regardless of whether harsh forms were excluded (β = .33 and p = .004 for limited, β = .34 and $p < .001$ for general), however the within-subject effects were more complex. There was not evidence for an effect of spanking on externalizing problems when using results from the general sample (β = .07 and .06, p = .224). In contrast, results indicated a *beneficial* effect on subsequent externalizing problems when excluding frequent spanking and the use of objects (β = −.17 and −.21, p = .018). Specifying the model as a typical CLPM yielded a reversed pattern of significance (p = .821 for limited, <.001 for general) with effect sizes both positive (β = .01 for limited, β = .12 for general), indicating adverse effects in the general sample, but no substantiative effects in the restricted sample.

DISCUSSION

In the present study, I aimed to contribute to the broader literature on spanking by using a method that allows for the examination of cross-lagged effects while controlling for between-subject differences. I applied an RI-CLPM to the results of Lansford et al. (2011, 2012), which both used the same dataset. Lansford et al. (2012) excluded children who were spanked with objects or at frequencies of once per week or higher, while Lansford et al. (2011) included these children. Previous within-subject results from Berry and Willoughby (2016) failed to find significant within-subject effects of spanking and more general physical punishment on externalizing, and I expected the same here for the general sample. However, there is a lack of studies using similar models on measures of exclusively mild spanking.

Results suggested an adverse effect of spanking on subsequent externalizing when using a typical CLPM on the general summary data from Lansford et al. (2011). However, consistent with my expectations and the results of Berry and Willoughby (2016), results were nonsignificant at a within-subject level. In other words, there was no substantiative evidence for an effect once between-subject differences were controlled for. It should be noted, however, that this does not necessarily imply that there is truly no effect, only that the present evidence is not significantly discordant with a lack of an effect. Such null results can be better interpreted through confidence intervals (CIs). CIs provide a range of values that are compatible with the data at a given confidence level. At a 95% confidence level, CIs ranged from medium-small adverse effects (βs = .173, .163) to very small beneficial ones (βs = −.040, −.038). Hence, large effects of either direction did not receive substantiative support.

In contrast to the results from the typical CLPM and recent meta-analyses using methods with the same issue, spanking was *inversely* associated with subsequent externalizing at a within-subject level when limited in frequency and excluding objects. This effect was small by conventional rule-of-thumb metrics, and about medium when compared to typical results for the field of psychology (Funder & Ozer, 2019). Some, however, have criticized general metrics like these for ignoring the context in which the effect occurs (e.g., Hill et al., 2008), suggesting that effects be interpreted contextually. Following this, effect sizes for the beneficial effect of limited spanking were larger than previously reported adverse cross-lagged effects of spanking and more general physical punishment (Ferguson, 2013; Larzelere et al., 2018). They were also somewhat larger than previously found cross-lagged effects of broader parenting dimensions on externalizing problems (Pinquart, 2017). Therefore, the effect of limited spanking was somewhat larger than is contextually typical, but not excessively so. It is worth noting that these comparisons are with results using cross-lagged models instead of with others examining within-subject relations, as few studies have focused on within-subject relations. Furthermore, while an effect of this magnitude is best supported by the data, 95% CIs suggest that beneficial effects anywhere from very small (βs = −.02, −.04) to large (βs = −.31, −.38) achieve close levels of support.

Within-subject results for both limited and general spanking differed from those that included between-subject differences, illustrating the importance of controlling for between-subject differences. In contrast to previous results that included between-subject differences, findings did not suggest that spanking causes increases in externalizing, and instead results suggested the opposite when only done infrequently and without objects. That is, although children who were spanked more frequently tended to have higher levels of externalizing problems, spanking did not appear to lead to subsequent heightened externalizing problems within individuals. Nonetheless, harsher spanking or other forms of physical punishment may still yield adverse effects, as these were not examined here individually.

Limitations

It is important to note the limitations of this study. While the RI-CLPM controls for raw group differences, it does not explicitly model developmental trends. The inclusion of time-specific means implicitly removes additive time effects that are globally present, but subject-specific trends may be needed. Note, however, that the inclusion of subject-specific trends could instead bias results if trend variance is a critical component of the process of interest rather than due to an underlying selection bias (Usami, Murayama, et al., 2019). Additionally, while the use of the RI-CLPM allows for the removal of trait-like confounders that are stable over time, this does not hold true for confounders that vary over time, but none were examined here.

Results from typical panel models, including the RI-CLPM used here, are dependent on the lag-length chosen. Long lag-lengths have been common in studies about spanking, but some have suggested that the optimal length for examining the effects of disciplinary actions is substantially shorter than lengths commonly used (Larzelere et al., 2010). Aside from simply using a different lag-length, some have proposed utilizing stochastic differential equations to model effects as a function of time, a method commonly referred to as continuous time modeling (cf. Oud & Jansen, 2000).

As noted by Ferguson (2013), effect sizes only represent an average effect, and differences across individual implementations may exist. Aside from by excluding spanking that is frequent or that involves the use of objects, no moderators were examined. Some previous studies have examined parental warmth or emotional support (e.g., McLoyd & Smith, 2004; Yildirim & Roopnarine, 2015) and more general constructs such as race and gender (e.g., Gunnoe & Mariner, 1997; MacKenzie et al., 2015). However, research examining moderators has generally been somewhat limited (Ferguson, 2013). Future studies would benefit from examining potential moderators on within-subject effects.

It should additionally be noted that this study focused only on externalizing problems, thus the same results do not necessarily hold true for internalizing problems or other outcomes. Unfortunately, few longitudinal studies have focused on the association with other outcomes (Ferguson, 2013), and even fewer when focusing on those with methodological strength. It should additionally be noted that the results here may not accurately represent the effectiveness of spanking at reducing specific behaviors, as to do such one would need to focus on the effect of spanking on the behaviors that each child was spanked for. Instead, this study focused on effects on externalizing behaviors in general, which is something that children are commonly spanked for but does not necessarily represent behavior-level effectiveness.

Given that this study simply reanalyzes previous results, all sample limitations of the previous studies apply. Measures of spanking were based upon maternal reports; however, paternal spanking is also relevant and thus total use is likely underreported. Within the subgroup of limited spanking, frequency and the use of objects were grouped together. As a result, the differences between forms noted here may not necessarily be from both, but could instead be from either aspect. There are also more aspects to harshness than frequency and the use of objects, and one could theoretically spank mildly with an object or severely with one's hand. Furthermore, it may be that higher-frequency spanking or the use of objects is confounded with other harsh or abusive practices rather than having true effect differences, as such practices were not examined. Moreover, given that subgroup comparisons effectively occur at a between-subject level, there may be selection bias on the basis of the subject-specific effects of spanking or even punishment in general. For instance, parents may spank with objects or more frequently when other tactics fail. Additionally, as the CDP is also only one source of data, future results could differ. It may have been preferable to reanalyze a study that originally found a more substantial link between limited spanking and adverse outcomes, as such appears to be the prevailing view in the literature. Unfortunately, few studies have examined explicitly limited usage at all, and it appears that no other published study has the benefits of Lansford et al. (2012).

Lastly, the differences between the full sample and restricted sample may be due to methodological differences within the original studies rather than a true difference in effect, and results should be interpreted with this in mind. While both studies used the same sample, it could be that more subtle differences in methodology caused differences in results.

CONCLUSIONS

Typical CLPMs assume a common mean across all individuals, yielding inaccurate results when group differences exist. Given that corrective actions are inherently confounded with what they intend to correct, group differences are practically inevitable and thus typical CLPMs are inappropriate for analyzing their effects. Further research utilizing improved methods is needed to examine results while avoiding such problems. Furthermore, as illustrated in the present study, different choices in how parents spank, such as in the use of an object or one's hand, may lead to differences in effects. Accordingly, research should examine different forms of spanking separately, at least in frequency and the use of objects.

Previous conclusions about spanking appear to have largely been based upon studies that include between-subject differences; however, results from such methods can be misleading. The present results suggest that previously reported adverse effects of spanking on externalizing problems are likely from confounding within-subject effects with between-subject associations rather than being indicative of causal effects. When done without objects and not used more frequently than about once per month, within-subject results instead indicated *beneficial* effects of spanking. Given these results, as well as prior ones (e.g., Berry & Willoughby, 2016; Larzelere et al., 2010, 2018), I suggest that policymakers and psychologists be cautious in interpreting results that include between-subject differences. Furthermore, care should be taken to distinguish results from different forms of physical punishment. Nonetheless, only externalizing problems were examined as an outcome, other outcomes could yield different results. Moreover, even if these findings held across samples and outcomes, it should be noted that non-empirical considerations may also be relevant in policy. Lastly, limited spanking may still carry risk

of escalation to harsher forms of physical punishment (Lansford et al., 2012). Spanking that was as frequent as about once per week or done with objects did not appear to be beneficial and could instead be adverse. Furthermore, such could lead to definite abuse.

ACKNOWLEDGMENTS

The author would like to thank those involved in the prior publications that made this study possible, as well as three anonymous reviewers for their comments. No funding was received for this study.

ORCID

Joshua Pritsker https://orcid.org/0000-0001-9647-6684

REFERENCES

Achenbach, T. M., & Edelbrock, C. S. (1986). *Manual for the teacher's report form and teacher version of child behavior profile*. University of Vermont, Department of Psychiatry.

Baumrind, D., Larzelere, R. E., & Cowan, P. A. (2002). Ordinary physical punishment: Is it harmful? Comment on Gershoff (2002). *Psychological Bulletin*, *128*, 580–589. https://doi.org/10.1037/0033-2909.128.4.580

Benjet, C., & Kazdin, A. E. (2003). Spanking children: The controversies, findings, and new directions. *Clinical Psychology Review*, *23*, 197–224. https://doi.org/10.1016/S0272-7358(02)00206-4

Berry, D., & Willoughby, M. T. (2016). On the practical interpretability of cross-lagged panel models: Rethinking a developmental workhorse. *Child Development*, *88*, 1186–1206. https://doi.org/10.1111/cdev.12660

Curran, P. J., Howard, A. L., Bainter, S., Lane, S. T., & McGinley, J. S. (2014). The separation of between-person and within-person components of individual change over time: A latent curve model with structured residuals. *Journal of Consulting and Clinical Psychology*, *82*, 879–894. https://doi.org/10.1037/a0035297

Dodge, K. A., Bates, J. E., & Pettit, G. S. (1990). Mechanisms in the cycle of violence. *Science*, *250*, 1678–1683. https://doi.org/10.1126/science.2270481

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

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*, 156–168. https://doi.org/10.1177/2515245919847202

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

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 & Adolescent Medicine*, *151*, 768. https://doi.org/10.1001/archpedi.1997.02170450018003

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. P. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods*, *20*, 102–116. https://doi.org/10.1037/a0038889

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*, 172–177. https://doi.org/10.1111/j.1750-8606.2008.00061.x

Hoffman, J., & Stawski, R. S. (2009). Persons as contexts: Evaluating between-person and within-person effects in longitudinal analysis. *Research in Human Development*, *6*, 97–120. https://doi.org/10.1080/15427600902911189

Hollingshead, A. B. (1975). *Four-factor index of social status* [Unpublished manuscript]. Yale University.

Lansford, J. E., Criss, M. M., Laird, R. D., Shaw, D. S., Pettit, G. S., Bates, J. E., & Dodge, K. A. (2011). Reciprocal relations between parents' physical discipline and children's externalizing behavior during middle childhood and adolescence. *Development and Psychopathology*, *23*, 225–238. https://doi.org/10.1017/s0954579410000751

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., & Baumrind, B. (2010). Are spanking injunctions scientifically supported? *Law and Contemporary Problems*, *73*, 57–87.

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., Gunnoe, M. L., & Ferguson, C. J. (2018). Improving causal inferences in meta-analyses of longitudinal studies: Spanking as an illustration. *Child Development*, *89*, 2038–2050. https://doi.org/10.1111/cdev.13097

Larzelere, R. E., Gunnoe, M. L., Ferguson, C. J., & Roberts, M. W. (2019). The insufficiency of the evidence used to categorically oppose spanking and its implications for families and psychological science: Comment on Gershoff et al (2018). *American Psychologist*, *74*, 497-499. https://doi.org/10.1037/amp0000461

MacKenzie, M. J., Nicklas, E., Brooks-Gunn, J., & Waldfogel, J. (2015). Spanking and children's externalizing behavior across the first decade of life: Evidence for transactional processes. *Journal of Youth and Adolescence*, *44*(3), 658–669. https://doi.org/10.1007/s10964-014-0114-y

McLoyd, V. C., & Smith, J. (2004). Physical discipline and behavior problems in African American, European American, and Hispanic children: Emotional support as a moderator. *Journal of Marriage and Family*, *64*, 40–53. https://doi.org/10.1111/j.1741-3737.2002.00040.x

Oud, J. H. L., & Jansen, R. A. R. G. (2000). Continuous time state space modeling of panel data by means of SEM. *Psychometrika*, *65*, 199–215. https://doi.org/10.1007/BF02294374

Pinquart, M. (2017). Associations of parenting dimensions and styles with externalizing problems of children and adolescents: An updated meta-analysis. *Developmental Psychology*, *53*, 873–932. https://doi.org/10.1037/dev0000295

R Core Team. (2019). *R: A language and environment for statistical computing*. https://www.r-project.org/

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36. https://doi.org/10.18637/jss.v048.i02

Smith, T. W., Davern, M., Freese, J., & Morgan, S. (2018). *General social surveys, 1972–2018*. NORC at the University Of Chicago. Data accessed from the GSS Data Explorer website at https://gssdataexplorer.norc.org

Usami, S., Murayama, K., & Hamaker, E. L. (2019). A unified framework of longitudinal models to examine reciprocal relations. *Psychological Methods*, *24*, 637–657. https://doi.org/10.1037/met0000210

Usami, S., Todo, N., & Murayama, K. (2019). Modeling reciprocal effects in medical research: Critical discussion on the current practices and potential alternative models. *PLoS One*, *14*. https://doi.org/10.1371/journal.pone.0209133

Yildirim, E. D., & Roopnarine, J. L. (2015). The mediating role of maternal warmth in the associations between harsh parental practices and externalizing and internalizing behaviors in Hispanic American, African American, and European American families.

Cultural Diversity and Ethnic Minority Psychology, *21*, 430–439. https://doi.org/10.1037/a0038210

SUPPORTING INFORMATION

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

How to cite this article: Pritsker, J. (2021). Spanking and externalizing problems: Examining within-subject associations. *Child Development*, 00, 1–8. https://doi.org/10.1111/cdev.13701