PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	A Multilevel Ecological Analysis of the Predictors of Spanking Across 65 Countries
AUTHORS	Ward, Kaitlin; Grogan-Kaylor, Andrew; Pace, Garrett; Cuartas, Jorge; Lee, Shawna

VERSION 1 – REVIEW

REVIEWER	Baba, Sachiko Osaka University, Bioethics and Public Policy, Department of Social Medicine, Graduate School of Medicine
REVIEW RETURNED	30-Nov-2020

GENERAL COMMENTS	Thank you for the opportunity to review this manuscript. This is a well-written manuscript regarding predictors of spanking using multilevel logistic regression analysis. My question is that whether you could consider adding the variable "country" as the macrolevel predictors. Some of the aspects in countries' differences were reflected by the macrolevel predictors, but others may not.

REVIEWER	Larzelere, Robert Oklahoma State University, Dept. of Human Devel. & Family Science
REVIEW RETURNED	24-Mar-2021

GENERAL COMMENTS	Some of the co-authors are well-known international leaders in the well-intentioned effort to protect children from physical violence, especially by opposing all physical punishment (“however light”) both in recommendations from their research and in supporting laws against physical punishment. Dr. Cuartas has done one of the first statistically controlled longitudinal studies outside of North America, which was also the first such study of physical punishment to document robustness across two ways of adjusting for initial differences on the outcome variable. The other co-authors have analyzed large datasets relevant to this issue, and this manuscript uses the largest dataset to date, as well as the most global in the number of countries involved. The study has other admirable features. Its huge sample size of people and of countries focuses on low and middle-income countries. It uses an excellent strategy for sampling respondents, especially for such a large survey. It one of the very few studies with a measure of spanking consistent with a common definition of spanking described in the first paragraph (open handed swat on the bottom). Most variables were missing fewer than 5% of the possible scores, and the data were scanned for outliers and multicollinearity. The analyses used multilevel modeling and showed it was warranted

by the ICC they reported. It is unclear whether they used the logistic equivalent of group-mean centering, which would make the micro-level effects pure within-country effects. They controlled for the quadratic effect of age as well as its linear effect.

With such a large sample size, it is important to distinguish between tiny and more meaningful associations, because tiny associations are almost always significant. Epidemiologists sometimes suggest that causal conclusions are more likely to be warranted when the odds ratio exceeds 2.0 (or, equivalently, less than .50), because residual confounding can often account for smaller associations. Cohen's guidelines for a small effect size of d = .20 would translate to OR = 1.44. Using either of those guidelines, parental beliefs that children need physical punishment is the only significant effect that meets those standards for a meaningful association. The significant OR's in Model 2 range from 1.01 to 1.24, with a median of 1.10 (equivalent to d = .053 or r = .026, which explains 0.07% of the variance in reported use of spanking. This illustrates the fact that null hypotheses are never actually true; failing to reject the null hypothesis merely indicates that the sample size was not large enough to detect the true effect size's difference from the null. That is why critiques of significance testing recommend emphasizing the effect size at least as much as its statistical significance – a recommendation that this paper needs to follow.

I have greater concerns, however, about the assumption underlying this study, that eliminating all disciplinary spanking is best for all children in all cultures. In medical practice, no established treatment would be totally eliminated from consideration until an alternative was shown to be consistently more effective for the same presenting problems that the established treatment had been used for (taking adverse side effects into account, too). I know of no research that has shown an alternative tactic that is more effective than spanking in the disciplinary situations for which cultures have considered it appropriate, with the same research methods on the same families.

I also know of no other issue in which current research claims that such a widespread traditional practice is now known to be so clearly harmful without exception that the evidence warrants imposing a zero tolerance of all spanking on all cultures worldwide. How adequate does the research evidence need to be for us affluent researchers of European descent to impose our parenting values on all parents everywhere? In my view, the current evidence falls so far short of adequate standards that it risks being counterproductive not only for the most at-risk children in less affluent families but also for the reputation of science.

A science skeptic could have a field day with this manuscript and related publications. Consider, for example, the statement at the end of the first paragraph: "rigorous evidence showing that spanking impairs children," citing Gershoff and Grogan-Kaylor's (2016) meta-analysis. The skeptic could quote from that very reference: "As most of the included studies were correlational or retrospective (72%), causal links between spanking and child outcomes cannot be established by these meta-analyses" (p. 464) and that they acknowledge that their strongest evidence from unadjusted longitudinal correlations "do not rule out the potential for a child elicitation effect; however, . . . longitudinal bivariate coefficients are decidedly stronger methodologically than within-time coefficients." (p.455). The skeptic might figure out from their Table 4 that 55% of

their evidence is from cross-sectional studies that cannot even ensure that the measure of physical punishment occurred before the child outcome. The skeptic might write an op-ed saying that if this is "rigorous evidence," it should be used to close all hospitals and intensive care units, because people treated in them have worse physical health than those outside hospitals at that time (cross-sectional correlations). As for the strongest causal evidence in this "rigorous evidence," longitudinal correlations would also show that those who were in intensive care units last year are more likely to have died since then than people not in ICU's. (The unadjusted longitudinal association of hospitalizations with subsequent physical health is r = -.33 (d = -.70: Angrist & Pischke, 2009, pp. 12-13), substantially larger than the r = .16 (d = .33) from your meta-analysis that the APA claimed to be sufficiently large to provide adequate causal evidence to oppose all spanking.)

I have become aware of this bias against all corrective actions because my research program has been searching for alternative disciplinary tactics that are more effective than spanking, which has resulted in our recognition that most non-randomized studies are biased against corrective actions, resulting in similar longitudinal evidence that they all look harmful, whether implemented by parents or professionals (e.g., psychotherapy, Ritalin, treatments for depression), even after controlling statistically for initial differences on the outcome variable. (You report that physical punishment has similar outcomes as ACE's; our replicated longitudinal studies show that physical punishment has similar outcomes as professional treatments for child behavioral problems. This is a pattern of data typical of all corrective actions: Larzelere, Lin, et al., 2018.) The skeptic could also discredit both the APA and the AAP because their evidence relied on your meta-analysis of unadjusted correlations, even though none of the four meta-analyses of physical punishment by other authors supported zero tolerance of all physical punishment, most of which excluded cross-sectional studies and did something to control for pre-existing differences on the child outcome. If this gets out (to the general public), it could discredit social science (and pediatricians?) not only in the area of physical discipline, but by association in any other areas of social scientific research that this skeptic disagrees with.

If the skeptic happens to be a feminist, she might be concerned about the huge increase in alleged rapes of children and adults in the 3 decades after Sweden banned spanking (e.g., a 73-fold increase in alleged rapes of children under 15, according to Swedish criminal statistics: Larzelere et al., 2013). She would be relieved to find your critique of this article by you and your co-authors in Holden et al. (2017), but could not find the increase in alleged rapes addressed specifically there (or anywhere else to my knowledge). But she looked up "the Bussmann et al. study [which] demonstrated only positive outcomes of banning physical punishment in Sweden" (Holden et al., p. 479). The feminist skeptic would be relieved to find from Bussmann et al. (2011) that Sweden has the lowest rate of "severe beating" of children. But she would be concerned about other comparisons, such as the finding that domestic partners are more likely to be insulted (79%) and tackled or hit (34%) than in countries without spanking bans (Spain and France: 35% & 18%, respectively). On the other hand, she would be glad that domestic partners are less likely to report being beaten or beaten up by their domestic partner in Sweden (average of 2 items: 4.5% vs. 13.4% in the two other spanking-ban countries (Germany & Austria) and 9.5%

in no-ban countries). Because less than 1/3 of Austrian and German parents were aware that their country's spanking ban applied to mild spanking, she might figure out that the study's Table 24.2 tests a natural experiment: Those who (incorrectly) thought that mild spanking was still legal were more likely to use it, but they were less likely to escalate to severe corporal punishment. This suggests that parents who can no longer use mild spanking are more likely to escalate to the point where they lash out in frustration with more severe corporal punishment, unless they are taught alternative tactics that are as effective as spanking when a child is at their most defiant.

This raises the other big issue from my perspective. To my knowledge, no one has shown that any alternative disciplinary tactic is more effective than nonabusive spanking when young children are very defiant. Timeout is an alternative emphasized in most of the empirically supported psychotherapies for treating oppositional defiance and conduct disorder in young children, yet you seem to support what you call "strong positive parenting," which means that it includes no negative disciplinary consequences imposed by parents at all, not even timeout. I only know of one study you have done on other disciplinary tactics, and that study failed to find any tactic that was correlated with significant reductions in externalizing or internalizing problems, despite investigating about 11 tactics with 88 analyses. You concluded that timeout and expressing disappointment to children were harmful, but the study did nothing to control for any confounding factors, not even pre-existing differences on the outcome variables.

I know of no other social work researchers who have done such an excellent job of introducing innovative statistical methods to improve causal inferences in non-randomized studies (e.g., fixed-effects regression, Bayesian methods). So it is a puzzle to me why such an advanced quantitative expert would ignore the ABC's of causal inferences when convenient (e.g., not requiring the cause to precede the effect in your "rigorous" meta-analysis, in your fixed-effects regression, and in your Bayesian methods (e.g., estimating the association of spanking during the previous month at age 5 with aggression during an indefinite recent time period at age 5, albeit controlling for Age-3 Aggression.). If our hypothetical science skeptic is as well read as you are, she might know that several statistical experts have shown that the more typical cross-lagged analyses are biased in the direction of stable between-person differences, including one article that shows that spanking has beneficial effects after analyzing the Fragile Families data "correctly" (in the Supplemental Material of Berry & Willoughby, 2017).

I do not question how well-intentioned you and your colleagues are. It would be ideal if children never had to experience any unnecessary pain. In a world in which few societal leaders seem to avoid stretching the truth if it helps them impose their values on everyone else, I expect science to resist that cultural tendency by striving to be as fair and objective as possible. Otherwise we social scientists are at risk of being guilty of facilitating the very kind of hegemony that influential leaders use to impose their values on less fortunate people regardless of their cultural and socioeconomic differences. In particular, we need causal evidence that is truly adequate to find alternatives to spanking that are effective in the disciplinary situations in which spanking has been considered most appropriate traditionally.

Miscellaneous details:

p. 4, 1st para.: I know of no evidence against spanking when defined as “physically hitting a child on the bottom with a bare hand to punish misbehavior.” In our latest meta-analysis, only one study limited spanking to that specific type (Lanford et al., 2012), and it was not associated significantly with any child outcome.

p. 4, 2nd para.: Our skeptic would find more problems with the references cited that spanking bans are associated with less community violence. Elgar et al. (2018) does not even ensure that the spanking ban preceded the measure of fighting among youth in those countries. At least for fighting among boys, the national differences are fully explained by stratifying the countries into (1) South Pacific Island countries, (2) Muslim-majority countries, and (3) other countries. After stratifying countries along these dimensions, the rate of fighting among boys is actually slightly higher in spanking-ban countries than in other countries, although not enough to be significant. (The means for girls’ fighting are still lower in spanking-ban countries than in other countries.)

p. 8: Since listwise deletion was used, what was the final sample size of people and countries in your analyses? For example, I assume that only 55 countries were used in Model 2?

p. 10: Was the comparison of MICS-4 vs. MICS-5 on the same countries, with a new sample of children of equivalent ages?

Table 2: Child age is in year, not months, which ought to be stated explicitly on p. 9, too.

Table 3: Why are the standard errors so large for HDI and Gender inequality? The effect size for HDI (OR = 1/29 = 3.45) is actually larger than for “Children need PP,” yet it is non-significant. Use three decimal places for SE and for the CI limits, at least to avoid SE = .00 and LCI = UCI. Why does the effect for the 2nd richest group switch sign after controlling for the macro-level predictors? In Model 1, the middle three wealth quintiles were significantly more likely to use spanking than either extreme. Education showed a similar curvilinear effect that persisted after controlling for macro-level predictors. Any explanation? Respondents who were female and who were biological parents were most likely to use spanking. Do they know something that our research is missing, perhaps some “mechanisms linking microsystem and macrosystem variables” (your p. 13) that current parenting research fails to take into account? We were surprised to find that mothers of toddlers reported using spanking out of concern for the welfare of their child, not for parent-oriented reasons (Lin et al., 2020).

Meta-Analyses of Physical Punishment (since Gershoff's 1st one in 2002):

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

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

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., & 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

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

Other References:

Angrist, J. D., & Pischke, J.-S. (2009). Mostly harmless econometrics: An empiricist's approach. Princeton University Press. https://doi.org/10.1515/9781400829828

Berry, D., & Willoughby, M. T. (2017). 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

Bussmann, K. D., Erthal, C., & Schroth, A. (2011). Effects of banning corporal punishment in Europe: A five-nation comparison. In J. E. Durrant & A. B. Smith (Eds.), Global pathways to abolish physical punishment: Realizing children's rights (pp. 299-322). Routledge

Elgar, F. J., Donnelly, P. D., Michaelson, V., Gariepy, G., Riehm, K. E., Walsh, S. D., & Pickett, W. (2018). Corporal punishment bans and physical fighting in adolescents: an ecological study of 88 countries. BMJ Open, 8(e021616), 1-8. https://doi.org/10.1136/bmjopen-2018-021616

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

Hamaker, E. L., Mulder, J. D., & van IJzendoorn, M. H. (2020, Dec). Description, prediction and causation: Methodological challenges of studying child and adolescent development. Developmental Cognitive Neuroscience, 46(100867). https://doi.org/10.1016/j.dcn.2020.100867

Holden, G. W., Grogan-Kaylor, A., Durrant, J. E., & Gershoff, E. T. (2017). Researchers deserve a better critique: Response to Larzelere, Gunnoe, Roberts, and Ferguson (2017). Marriage & Family Review, 53, 465-490. https://doi.org/10.1080/01494929.2017.1308899

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., Lin, H., Payton, M. E., & Washburn, I. J. (2018). Longitudinal biases against corrective actions. Archives of Scientific Psychology, 6, 243-250. https://doi.org/10.1037/arc0000052 Larzelere, R. E., Swindle, T., & Johnson, B. R. (2013). Swedish trends in criminal assaults against minors since banning spanking, 1981-2010. International Journal of Criminology and Sociology, 2, 129-137. https://doi.org/http://www.lifescienceglobal.com/media/zj_fileseller/files/IJCSV2A13-Larzelere.pdf Lin, H., Ritchie, K. L., & Larzelere, R. E. (2020). Applying a momentary parenting goal-regulation model to discipline episodes with toddlers. Journal of Child and Family Studies, 29(4), 1055-1069. https://doi.org/10.1007/s10826-020-01698-1

VERSION 1 – AUTHOR RESPONSE

Reviewer 1

Thank you for the opportunity to review this manuscript. This is a well-written manuscript regarding predictors of spanking using multilevel logistic regression analysis. My question is that whether you could consider adding the variable "country" as the macrolevel predictors. Some of the aspects in countries' differences were reflected by the macrolevel predictors, but others may not.

Response: We thank the reviewer for this comment. Per standard practice in the multi-country child development literature, we have chosen to add country as a *random intercept* to our multilevel models. Adding country as an indicator variable would use too many degrees of freedom, and would eliminate our ability to explore country level effects.

References:

Allison, P. (2009). Fixed Effects Regression Models. Thousand Oaks, CA: SAGE Publications Inc. https://doi.org/10.4135/9781412993869

Kreft, I., & de Leeuw, J. (1998). Introducing Multilevel Modeling. London, UK: SAGE Publications. https://doi.org/10.4135/9781849209366

Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical linear models: applications and data analysis methods. Thousand Oaks, CA: Sage Publications.

Preface to Reviewer 2: We would like to thank this reviewer for their detailed comments and insights. Many of the reviewer's suggestions relate to critiques of other manuscripts. While these critiques have stimulated much thought and discussion among the co-authors, we feel that many go beyond what can be addressed in an R&R. We also believe that some of the reviewer's comments do not reflect the larger global conversation surrounding corporal punishment, which has been a truly international one, including a United Nations supported study of violence against children, which recommended that parents no longer use corporal punishment. Further, the United Nations Study of Violence Against Children has been endorsed by a resolution of the General Assembly of the United Nations, which specifically called for an end to violent discipline against children. Therefore, we will not be

following the reviewer's recommended approach to reframing the study. However, we have wholeheartedly attempted to address the reviewer's methodological concerns with the manuscript.

We also sought out guidance from the editor, who made specific recommendations as to which of the reviewer's points to address. We will be addressing the concerns that the editor, as well as our team, felt would strengthen the manuscript.

References:

Pinheiro, P. S., & United Nations Secretary General's Study on Violence Against, C. (2006). World report on violence against children [electronic resource]. United Nations. http://www.violencestudy.org/a553

UN General Assembly A/RES/62/141

Reviewer 2

Some of the co-authors are well-known international leaders in the well-intentioned effort to protect children from physical violence, especially by opposing all physical punishment ("however light") both in recommendations from their research and in supporting laws against physical punishment. Dr. Cuartas has done one of the first statistically controlled longitudinal studies outside of North America, which was also the first such study of physical punishment to document robustness across two ways of adjusting for initial differences on the outcome variable. The other co-authors have analyzed large datasets relevant to this issue, and this manuscript uses the largest dataset to date, as well as the most global in the number of countries involved.

The study has other admirable features. Its huge sample size of people and of countries focuses on low and middle-income countries. It uses an excellent strategy for sampling respondents, especially for such a large survey. It one of the very few studies with a measure of spanking consistent with a common definition of spanking described in the first paragraph (open handed swat on the bottom). Most variables were missing fewer than 5% of the possible scores, and the data were scanned for outliers and multicollinearity. The analyses used multilevel modeling and showed it was warranted by the ICC they reported. It is unclear whether they used the logistic equivalent of group-mean centering, which would make the micro-level effects pure within-country effects. They controlled for the quadratic effect of age as well as its linear effect.

Response: We thank the reviewer for highlighting these strengths of our study. We wish to clarify that we did not employ group mean centering in this study.

With such a large sample size, it is important to distinguish between tiny and more meaningful associations, because tiny associations are almost always significant. Epidemiologists sometimes suggest that causal conclusions are more likely to be warranted when the odds ratio exceeds 2.0 (or, equivalently, less than .50), because residual confounding can often account for smaller associations. Cohen's guidelines for a small effect size of d = .20 would translate to OR = 1.44. Using either of those guidelines, parental beliefs that children need physical punishment is the only significant effect that meets those standards for a meaningful association. The significant OR's in Model 2 range from 1.01 to 1.24, with a median of 1.10 (equivalent to d = .053 or r = .026, which explains 0.07% of the variance in reported use of spanking. This illustrates the fact that null hypotheses are never actually true; failing to reject the null hypothesis merely indicates that the sample size was not large enough to detect the true effect size's difference from the null. That is why critiques of significance testing

recommend emphasizing the effect size at least as much as its statistical significance – a recommendation that this paper needs to follow.

Response: We would like to note that our discussion section primarily focuses on the outcomes with the larger effect sizes. To further address the reviewer's point, we have now added additional information to the Limitations section of our manuscript, and have provided justification why even smaller effect sizes found in the manuscript may still merit consideration for future research.

Reviewer 2, continued:

I have greater concerns, however, about the assumption underlying this study, that eliminating all disciplinary spanking is best for all children in all cultures. In medical practice, no established treatment would be totally eliminated from consideration until an alternative was shown to be consistently more effective for the same presenting problems that the established treatment had been used for (taking adverse side effects into account, too). I know of no research that has shown an alternative tactic that is more effective than spanking in the disciplinary situations for which cultures have considered it appropriate, with the same research methods on the same families.

I also know of no other issue in which current research claims that such a widespread traditional practice is now known to be so clearly harmful without exception that the evidence warrants imposing a zero tolerance of all spanking on all cultures worldwide. How adequate does the research evidence need to be for us affluent researchers of European descent to impose our parenting values on all parents everywhere? In my view, the current evidence falls so far short of adequate standards that it risks being counterproductive not only for the most at-risk children in less affluent families but also for the reputation of science.

A science skeptic could have a field day with this manuscript and related publications. Consider, for example, the statement at the end of the first paragraph: "rigorous evidence showing that spanking impairs children," citing Gershoff and Grogan-Kaylor's (2016) meta-analysis. The skeptic could quote from that very reference: "As most of the included studies were correlational or retrospective (72%), causal links between spanking and child outcomes cannot be established by these meta-analyses" (p. 464) and that they acknowledge that their strongest evidence from unadjusted longitudinal correlations "do not rule out the potential for a child elicitation effect; however, . . . longitudinal bivariate coefficients are decidedly stronger methodologically than within-time coefficients." (p.455). The skeptic might figure out from their Table 4 that 55% of their evidence is from cross-sectional studies that cannot even ensure that the measure of physical punishment occurred before the child outcome. The skeptic might write an op-ed saying that if this is "rigorous evidence," it should be used to close all hospitals and intensive care units, because people treated in them have worse physical health than those outside hospitals at that time (cross-sectional correlations). As for the strongest causal evidence in this "rigorous evidence," longitudinal correlations would also show that those who were in intensive care units last year are more likely to have died since then than people not in ICU's. (The unadjusted longitudinal association of hospitalizations with subsequent physical health is r = -.33 (d = -.70: Angrist & Pischke, 2009, pp. 12-13), substantially larger than the r = .16 (d = .33) from your meta-analysis that the APA claimed to be sufficiently large to provide adequate causal evidence to oppose all spanking.)

Response: We agree with the reviewer's contention that much of the evidence in the Gershoff and Grogan-Kaylor (2016) meta-analysis is from cross-sectional studies. However, the evidence from this meta-analysis of 50 years of empirical research on corporal punishment needs to be weighed *in conjunction with* a number of recent important studies that employ careful statistical analysis to make

stronger causal conclusions about the effects of spanking and find that the association of spanking with child behavior problems persists despite the strong statistical controls that have been employed. We have added a few of these studies as citations to support our assertion that spanking harms children.

References:

Gershoff, E. T., Sattler, K. M. P., & Ansari, A. (2018). Strengthening Causal Estimates for Links Between Spanking and Children's Externalizing Behavior Problems. Psychological Science, 29(1), 110–120. https://doi.org/10.1177/0956797617729816

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(3), 838–843. https://doi.org/10.1111/j.1467-8624.2011.01732.x

Cuartas, J., McCoy, D. C., Grogan-Kaylor, A., & Gershoff, E. (2020). Physical Punishment as a Predictor of Early Cognitive Development: Evidence From Econometric Approaches. Developmental Psychology. https://doi.org/10.1037/dev0001114

Ma, J., Grogan-Kaylor, A., & Lee, S. J. (2018). Associations of neighborhood disorganization and maternal spanking with children's aggression: A fixed-effects regression analysis. Child Abuse and Neglect, 76, 106–116. https://doi.org/10.1016/j.chiabu.2017.10.013

Reviewer 2, continued:

I have become aware of this bias against all corrective actions because my research program has been searching for alternative disciplinary tactics that are more effective than spanking, which has resulted in our recognition that most non-randomized studies are biased against corrective actions, resulting in similar longitudinal evidence that they all look harmful, whether implemented by parents or professionals (e.g., psychotherapy, Ritalin, treatments for depression), even after controlling statistically for initial differences on the outcome variable. (You report that physical punishment has similar outcomes as ACE's; our replicated longitudinal studies show that physical punishment has similar outcomes as professional treatments for child behavioral problems. This is a pattern of data typical of all corrective actions: Larzelere, Lin, et al., 2018.) The skeptic could also discredit both the APA and the AAP because their evidence relied on your meta-analysis of unadjusted correlations, even though none of the four meta-analyses of physical punishment by other authors supported zero tolerance of all physical punishment, most of which excluded cross-sectional studies and did something to control for pre-existing differences on the child outcome. If this gets out (to the general public), it could discredit social science (and pediatricians?) not only in the area of physical discipline, but by association in any other areas of social scientific research that this skeptic disagrees with.

If the skeptic happens to be a feminist, she might be concerned about the huge increase in alleged rapes of children and adults in the 3 decades after Sweden banned spanking (e.g., a 73-fold increase in alleged rapes of children under 15, according to Swedish criminal statistics: Larzelere et al., 2013). She would be relieved to find your critique of this article by you and your co-authors in Holden et al. (2017), but could not find the increase in alleged rapes addressed specifically there (or anywhere else to my knowledge). But she looked up "the Bussmann et al. study [which] demonstrated only positive outcomes of banning physical punishment in Sweden" (Holden et al., p. 479). The feminist skeptic would be relieved to find from Bussmann et al. (2011) that Sweden has the lowest rate of "severe

beating" of children. But she would be concerned about other comparisons, such as the finding that domestic partners are more likely to be insulted (79%) and tackled or hit (34%) than in countries without spanking bans (Spain and France: 35% & 18%, respectively). On the other hand, she would be glad that domestic partners are less likely to report being beaten or beaten up by their domestic partner in Sweden (average of 2 items: 4.5% vs. 13.4% in the two other spanking-ban countries (Germany & Austria) and 9.5% in no-ban countries). Because less than 1/3 of Austrian and German parents were aware that their country's spanking ban applied to mild spanking, she might figure out that the study's Table 24.2 tests a natural experiment: Those who (incorrectly) thought that mild spanking was still legal were more likely to use it, but they were less likely to escalate to severe corporal punishment. This suggests that parents who can no longer use mild spanking are more likely to escalate to the point where they lash out in frustration with more severe corporal punishment, unless they are taught alternative tactics that are as effective as spanking when a child is at their most defiant.

This raises the other big issue from my perspective. To my knowledge, no one has shown that any alternative disciplinary tactic is more effective than nonabusive spanking when young children are very defiant. Timeout is an alternative emphasized in most of the empirically supported psychotherapies for treating oppositional defiance and conduct disorder in young children, yet you seem to support what you call "strong positive parenting," which means that it includes no negative disciplinary consequences imposed by parents at all, not even timeout. I only know of one study you have done on other disciplinary tactics, and that study failed to find any tactic that was correlated with significant reductions in externalizing or internalizing problems, despite investigating about 11 tactics with 88 analyses. You concluded that timeout and expressing disappointment to children were harmful, but the study did nothing to control for any confounding factors, not even pre-existing differences on the outcome variables.

I know of no other social work researchers who have done such an excellent job of introducing innovative statistical methods to improve causal inferences in non-randomized studies (e.g., fixed-effects regression, Bayesian methods). So it is a puzzle to me why such an advanced quantitative expert would ignore the ABC's of causal inferences when convenient (e.g., not requiring the cause to precede the effect in your "rigorous" meta-analysis, in your fixed-effects regression, and in your Bayesian methods (e.g., estimating the association of spanking during the previous month at age 5 with aggression during an indefinite recent time period at age 5, albeit controlling for Age-3 Aggression.). If our hypothetical science skeptic is as well read as you are, she might know that several statistical experts have shown that the more typical cross-lagged analyses are biased in the direction of stable between-person differences, including one article that shows that spanking has beneficial effects after analyzing the Fragile Families data "correctly" (in the Supplemental Material of Berry & Willoughby, 2017).

I do not question how well-intentioned you and your colleagues are. It would be ideal if children never had to experience any unnecessary pain. In a world in which few societal leaders seem to avoid stretching the truth if it helps them impose their values on everyone else, I expect science to resist that cultural tendency by striving to be as fair and objective as possible. Otherwise we social scientists are at risk of being guilty of facilitating the very kind of hegemony that influential leaders use to impose their values on less fortunate people regardless of their cultural and socioeconomic differences. In particular, we need causal evidence that is truly adequate to find alternatives to spanking that are effective in the disciplinary situations in which spanking has been considered most appropriate traditionally.

Miscellaneous details:

p. 4, 1st para.: I know of no evidence against spanking when defined as “physically hitting a child on the bottom with a bare hand to punish misbehavior.” In our latest meta-analysis, only one study limited spanking to that specific type (Lanford et al., 2012), and it was not associated significantly with any child outcome.

Response: We have removed “to punish misbehavior” from the definition on page 4 to better capture varying definitions of spanking. The Gershoff and Grogan-Kaylor (2016) meta-analysis includes a discussion of definitions of spanking across studies.

p. 4, 2nd para.: Our skeptic would find more problems with the references cited that spanking bans are associated with less community violence. Elgar et al. (2018) does not even ensure that the spanking ban preceded the measure of fighting among youth in those countries. At least for fighting among boys, the national differences are fully explained by stratifying the countries into (1) South Pacific Island countries, (2) Muslim-majority countries, and (3) other countries. After stratifying countries along these dimensions, the rate of fighting among boys is actually slightly higher in spanking-ban countries than in other countries, although not enough to be significant. (The means for girls’ fighting are still lower in spanking-ban countries than in other countries.)

Response: The reviewer is correct that Elgar et al. (2018) did not account for the timing of bans in their analysis. Nevertheless, Elgar et al.’s paper is among the most rigorous studies to-date on this topic. To our knowledge, there is not a Elgar et al. replication paper or an extension of their analysis that accounts for the timing of bans or involves the stratification of countries, so we are unable to incorporate this level of detail in the manuscript. We have tempered the language of this sentence to address the reviewer’s concern.

p. 8: Since listwise deletion was used, what was the final sample size of people and countries in your analyses? For example, I assume that only 55 countries were used in Model 2?

Response: This information is included in Table 3. Specifically, see the note to Table 3 that includes the sample size and country number included in each model. We have also added this detail to the Analytic Strategy section.

p. 10: Was the comparison of MICS-4 vs. MICS-5 on the same countries, with a new sample of children of equivalent ages?

Response: The comparison of MICS4 and MICS5 was with a new sample of children of nearly equivalent ages. The fourth round includes children ages 2-14 years and the fifth round includes children ages 1-14 years. Of the 55 countries included in the comparison in Model 2, 25 only participated in the fourth round, 23 only participated in the fifth round, and 7 participated in both rounds. At the end of the Results section, we now explain that the lower odds of spanking at round 5 may be due to 3 reasons: 1) the inclusion of 1-year-olds in the fifth round, 2) differences between countries since only 7 countries participated in both rounds, and 3) the prevalence of spanking may also be declining over time.

Table 2: Child age is in year, not months, which ought to be stated explicitly on p. 9, too.

Response: We have changed “months” to “years” in Table 2. In the first paragraph of the Results, we have added further clarification that age is measured in years.

Table 3: Why are the standard errors so large for HDI and Gender inequality? The effect size for HDI (OR = 1/29 = 3.45) is actually larger than for “Children need PP,” yet it is non-significant. Use three decimal places for SE and for the CI limits, at least to avoid SE = .00 and LCI = UCI. Why does the effect for the 2nd richest group switch sign after controlling for the macro-level predictors? In Model 1, the middle three wealth quintiles were significantly more likely to use spanking than either extreme. Education showed a similar curvilinear effect that persisted after controlling for macro-level predictors. Any explanation? Respondents who were female and who were biological parents were most likely to use spanking. Do they know something that our research is missing, perhaps some “mechanisms linking microsystem and macrosystem variables” (your p. 13) that current parenting research fails to take into account? We were surprised to find that mothers of toddlers reported using spanking out of concern for the welfare of their child, not for parent-oriented reasons (Lin et al., 2020).

Response: Although we may not be able to discern the exact reason why the standard errors of HDI and gender inequality are so large, it is possible that 1) HDI and gender inequality are simply poor predictors of spanking when controlling for other factors, or 2) there may be low coverage in one of the cells to provide precise estimates (e.g., it may be very unlikely for countries with low gender inequality to have high rates of spanking). We have followed the reviewer’s recommendation to expand the estimates in Table 3 to three decimal places. We have also added a sentence to the Limitations section regarding the directionality of the relationship between wealth quintiles and spanking switching after other variables were taken into account.