HHS Public Access

Author manuscript

Dev Psychol. Author manuscript; available in PMC 2021 March 22.

Published in final edited form as:

Dev Psychol. 2020 November ; 56(11): 2013–2026. doi:10.1037/dev0001114.

Physical Punishment as a Predictor of Early Cognitive Development: Evidence From Econometric Approaches

Jorge Cuartas,
Harvard Graduate School of Education, Harvard University

Dana Charles McCoy,
Harvard Graduate School of Education, Harvard University

Andrew Grogan-Kaylor,
School of Social Work, University of Michigan

Elizabeth Gershoff
Department of Human Development and Family Sciences and the Population Research Center, University of Texas at Austin.

Abstract

This study estimates the effect of physical punishment on the cognitive development of 1,167 low-income Colombian children (M_{age} = 17.8 months old) using 3 analytic strategies: lagged-dependent variables, a difference-in-differences-like approach (DD), and a novel strategy combining matching with a DD-like approach. Across approaches, physical punishment at ages 9–26 months predicted reductions in children's cognitive development of 0.08–0.21 *SD* at ages 27–46 months. These results, plus null results of falsification tests, strengthen the argument that physical punishment leads to slower cognitive growth and illustrate the utility of alternative statistical methods to reduce problems of selection bias in developmental research.

Keywords

physical punishment; early childhood development; child discipline; corporal punishment; Colombia

Developmental science has long been invested in understanding the ways in which environmental characteristics may affect children's outcomes. At the same time, establishing causality in developmental research remains challenging because human development is shaped by interactions between biological and ecological factors (Bronfenbrenner & Morris, 2007; Sameroff, 2010) and because potential predictors of children's development often co-

Correspondence concerning this article should be addressed to Jorge Cuartas, Harvard Graduate School of Education, Harvard University, 14 Appian Way, Cambridge, MA 02138. jcuartas@g.harvard.edu.

We thank Eric Taylor, Jon Star, Catalina Rey-Guerra, and the members of the SEED lab at Harvard University for their useful comments. All data for this study were taken from publicly available data sources. In particular, individual and household data are available at http://data-archive.ac.uk/ (Fitzsimons, Rubio-Codina, Andrew, & Attanasio, 2010), and municipality data are available from Universidad de los Andes' Department of Economics Data Center at https://datoscede.uniandes.edu.co (Acevedo & Bornacelly, 2014).

occur, making it difficult to isolate the specific effect of any one predictor (Foster, 2010; Miller, Henry, & Votruba-Drzal, 2016). Whereas other fields have addressed this challenge of "selection" through experimental manipulation, in developmental research random assignment of children to different contextual conditions (e.g., parental practices, environmental hazards) is often impractical or unethical. Developmental psychologists, then, must largely rely on alternative methods for improving the internal validity of their estimates and discarding plausible alternative explanations (Duncan, Magnuson, & Ludwig, 2004; Foster, 2010; Miller et al., 2016).

Within this context, research on the effects of physical punishment (i.e., using physical force with the objective to cause pain or discomfort to correct or punish a child's behavior; Gershoff, 2002) on children's well-being constitutes an interesting case that has been debated within the field (e.g., Berry & Willoughby, 2017; Gershoff, 2013; Gershoff, Goodman, et al., 2018; Larzelere, Gunnoe, & Ferguson, 2018). On the one hand, multiple developmental theories support the claim that physical punishment is detrimental for children's development (Gershoff, 2002) and a large body of evidence shows systematic associations between physical punishment and deleterious child outcomes (Gershoff & Grogan-Kaylor, 2016). On the other hand, questions have been raised regarding whether these associations are causal in nature (e.g., see Larzelere et al., 2018), mainly due to concerns around reverse-causality (or simultaneous causality) and of the need for more (a) studies that account for key potential confounding characteristics, including risk factors that may co-occur with physical punishment such as socioeconomic disadvantage; (b) within-study replications or robustness checks (e.g., employing different methods and finding consistent results); and (c) falsification tests to assess the validity of the findings. Despite these arguments, the plausibility of the underlying assumptions to identify a causal effect (also known as *identifying assumptions*) have been rarely discussed in the physical punishment literature specifically, or in the developmental literature more broadly, even though these assumptions are central to assessing the plausibility of a causal link (Angrist & Pischke, 2011).

The purpose of this study was to contribute to this conversation by systematically testing the plausibility of causal links between children's exposure to physical punishment and cognitive skills during early childhood. To do so, we employed three empirical approaches with relatively weaker (i.e., more plausible) identifying assumptions than those used in previous research. These three methods constituted within-study replications to test the extent to which our results were consistent or robust across different methods (Duncan, Engel, Claessens, & Dowsett, 2014).

Theoretical and Empirical Links Between Physical Punishment and Cognitive Development

Two meta-analyses have addressed the question of whether physical punishment by parents is linked with children's cognitive development. The first meta-analysis (Ferguson, 2013) included only longitudinal studies and found that physical punishment predicts slower growth in cognitive skills over time, $d = -.11$ (95% CI: −.05/−17). A second meta-analysis

that included cross-sectional and longitudinal studies (Gershoff & Grogan-Kaylor, 2016) also found consistent associations between physical punishment and lower cognitive ability, $d = -.17$ (95% CI: −.01/−.32). As an illustration, one of the studies included in both meta-analyses (Berlin et al., 2009) used cross-lagged panel analysis to find that children's exposure to spanking at age one predicted slower growth in cognitive development by age three (notably using the same measure of cognitive development used in the current study). One study found that only physically abusive methods (e.g., hitting with a fist, choking, beating), and not mild physical punishment (e.g., spanking, slapping), were linked with lower math and preliteracy scores (Font & Cage, 2018), leaving open the question of what degree of physical force used against children triggers changes that affect cognitive development. Yet there is a consistent body of research finding that physical punishment and physical abuse are linked with the same child outcomes, just to a different degree (Gershoff & Grogan-Kaylor, 2016), and that physical punishment is linked with psychological distress and mental health problems even after controlling for childhood exposure to physical abuse (Afifi et al., 2017). The balance of studies thus points to physical punishment as a potential sufficient cause of slower growth in children's cognitive skills over time.

Why would physical punishment be linked with children's cognitive functioning? The primary candidate is changes to children's brain structures and functioning, particularly in response to threat. Physical punishment is understood to be a major source of stress for children (Gershoff, 2016). Physical punishment meets the criteria of a toxic stressor because the punishment causes pain and distress to the child and because the parent or caregiver, who should be a source of comfort and support, is the source of stress itself (Gershoff, 2016; Shonkoff, Boyce, & McEwen, 2009). The repeated activation of children's stress-response systems every time they are physically punished can result in allostatic load, which can in turn can lead to changes in the structure and functioning of the prefrontal cortex, the amygdala, and the hippocampus (Danese & McEwen, 2012).

Few studies have tested this possibility explicitly to provide support for the hypothesis that exposure to physical punishment results in activation of the stress-response system and in long-term changes to the brain. In one such study, infants who were physically punished were found to exhibit high cortisol reactivity to stress, which implies that the stress response system (namely the hypothalamic-pituitary-adrenal axis) may be a link between physical punishment and cognitive deficits in children (Bugental, Martorell, & Barraza, 2003). Additional support for the notion that physical punishment may have direct effects on the brain comes from findings that young adults with histories of chronic and harsh physical punishment have less gray matter in their prefrontal cortex than do peers who did not experience it (Tomoda et al., 2009) and have evidence of physical alterations in regions of the brain related to memory (Sheu, Polcari, Anderson, & Teicher, 2010).

In theory, physical punishment may exert a negative influence on children even in the context of exposure to other contextual stressors. A large body of evidence suggests that punishment works similarly to other adverse childhood experiences, having additive negative consequences for children beyond other adverse exposures (Afifi et al., 2017). Moreover, it is likely that the effects of physical punishment may be even more severe (not just additive, but potentially multiplicative) in the context of other sources of adversity. Indeed,

neurobiological models of human development indicate that parental warmth, responsiveness, and protection are fundamental to buffer the harmful effects of contextual stressors (Gunnar, Hostinar, Sanchez, Tottenham, & Sullivan, 2015). Consequently, the use of corporal punishment (instead of parental practices that promote a sense of safety and comfort) might actually exacerbate the negative consequences of other adverse exposures. As such, exploring the effects of physical punishment in contexts with particularly high levels of baseline adversity is a needed area of research.

Issues of Causality

Despite the consistency of the research to date linking physical punishment to lower cognitive ability, recent studies (e.g., Berry & Willoughby, 2017; Larzelere et al., 2018) have pointed out that the methods used to analyze the effects of physical punishment on children's developmental outcomes generally suffer from important limitations that make it difficult to infer credible causal conclusions. For instance, although there is a well-established association between physical punishment and children's externalizing behavior problems (Gershoff & Grogan-Kaylor, 2016), these findings could suffer from *reverse causation*, such that children's behavioral problems (and other skills and behaviors, more broadly) could also elicit more physical punishment from parents over time (Larzelere et al., 2018). That said, reverse causation does not automatically rule out the hypothesized causation because it is possible for there to be simultaneous causality. Indeed, several studies employing cross-lagged panel models have found evidence for *simultaneous causality*, where spanking predicts more behavior problems over time and behavior problems elicit more spanking over time (Berlin et al., 2009; Maguire-Jack, Gromoske, & Berger, 2012; Gershoff, Lansford, Sexton, Davis-Kean, & Sameroff, 2012; Wang & Kenny, 2014). Although such a coercive cycle is thought to explain the link between physical punishment and children's behavior problems (Patterson, 1982), less evidence exists with regard children's cognitive skills.

A perhaps more concerning threat to causal inference in the physical punishment literature is the phenomenon known as *omitted variable bias*. Omitted variable bias (which is also known as selection, confounding, endogeneity bias, or the third variable problem) occurs when unobserved characteristics that underlie both parents' use of physical punishment and children's development are not accounted for in the analysis (Duncan et al., 2004). For example, even though most studies control for potential confounding characteristics such as household socioeconomic status or parental education, other less easily measurable characteristics such as parents' and children's shared genes or community norms around child rearing may simultaneously influence both parents' use of physical punishment and children's development. Indeed, one meta-analysis found that the strength of the negative associations between physical punishment and child outcomes are sensitive (i.e., not robust) to different model specifications and covariates (Larzelere et al., 2018), although another meta-analysis instead found the associations to be robust to a range of covariates (Gershoff & Grogan-Kaylor, 2016).

These concerns about the physical punishment literature beg the question: Under nonexperimental conditions, what constitutes credible evidence for causality? In general, the

strength of such evidence will depend on the plausibility of the underlying identifying assumptions, which are largely untestable. When attempting to establish causality, all assumptions must be made explicit, and stronger assumptions will demand more information to make a convincing case for a causal link (Foster, 2010). As such, weaker (i.e., more plausible) assumptions, the robustness of results to alternative approaches, and results that align with theory each make a causal link more credible (Angrist & Pischke, 2011; Duncan et al., 2014; Duncan et al., 2004; Foster, 2010; Miller et al., 2016).

A few recent studies on the effects of physical punishment have used alternative approaches aimed at mitigating selection or omitted variable bias in order to draw more convincing causal conclusions. For example, Ma, Grogan-Kaylor, and Lee (2018) found associations between maternal spanking during early childhood and higher levels of child aggression employing a fixed-effects model, which is a method that evaluates changes over time withinindividuals, using each individual as his or her own statistical control, in order to account for time-invariant characteristics (similarly to difference-in-differences methods; Angrist & Pischke, 2011). Two additional studies used an approach known as propensity score matching (PSM) that constructs an artificial control group (e.g., children not exposed to physical punishment) who are similar to children exposed to physical punishment on a number of relevant, observed characteristics as a means of mimicking the conditions produced by random assignment (Caliendo & Kopeinig, 2008). For example, using PSM in a nationally representative sample of over 12,000 families, Gershoff, Sattler, and Ansari (2018) found that children spanked at age 5 experienced increases in their externalizing problem behaviors by ages 6 and 8 years compared with a group of children not exposed to physical punishment but otherwise equal (i.e., matched) on a set of 38 individual, family, and cultural characteristics. A study in Japan of more than 29,000 children also used PSM and found that spanked children exhibited higher increases in behavior problems from age 3.5 to age 5.5 than did nonspanked children even after they were matched on 28 characteristics (Okuzono, Fujiwara, Kato, & Kawachi, 2017).

Despite the above evidence, individual fixed-effects models (i.e., within-person analyses) and matching techniques rely on relatively strong—and therefore potentially implausible—identifying assumptions. The internal validity of the method of fixed effects relies on the assumption that omitted characteristics that simultaneously affect the use of physical punishment and children's outcomes do not vary across time (Angrist & Pischke, 2011; Blundell & Dias, 2009). As such, if a variable omitted from the model is likely to change over time and also jointly influences parental use of physical punishment and children's outcomes (e.g., parental depression or social norms around child rearing), such a variable will threaten validity. Matching techniques such as PSM rely on the same assumption that underlies linear regression models, namely *conditional independence*, which implies that any omitted or unobserved characteristic correlated simultaneously with the use of physical punishment and children's outcomes will bias the results (Caliendo & Kopeinig, 2008). For instance, not including children's exposure to community violence, which has been shown to correlate both with parental use of physical punishment (Coulton, Crampton, Irwin, Spilsbury, & Korbin, 2007) and with children's developmental outcomes (Horn & Trickett, 1998), will constitute a threat to validity. The same will be true for several other proximal

and distal factors that are likely to simultaneously influence parental use of physical punishment and children's development (Bronfenbrenner & Morris, 2007; McCoy, 2013).

In addition to relying on relatively strong identifying assumptions, previous studies concerning the association of physical punishment with children's cognitive development have at least three additional limitations. First, infrequent use of robustness and falsification checks has raised concerns about the validity of prior studies (Larzelere et al., 2018) and this in turn raises questions about the sensitivity of past results to different methodologies or specifications (Duncan et al., 2014; Foster, 2010). Second, the association between physical punishment and children's cognitive development has been relatively understudied in comparison to its association with children's behavioral outcomes (see Ferguson, 2013; Gershoff & Grogan-Kaylor, 2016). Finally, most research has been conducted with U.S. samples (Gershoff & Grogan-Kaylor, 2016). An increasing number of studies around the world have found that physical punishment is linked with detrimental child outcomes (e.g., 8 countries: Alampay et al., 2017; 62 countries: Pace, Lee, & Grogan-Kaylor, 2019), yet these studies have largely not used strategies designed to improve causal inferences.

The Present Study

The primary objective of this study was to examine the plausibility of a causal effect of physical punishment on cognitive development during early childhood. To do so, we responded to several key suggestions from previous literature. First, we included in our analyses variables that may mitigate omitted variable bias (Larzelere et al., 2018), including child characteristics, such as their initial level of cognitive skill, and characteristics of the parent, such as their parenting beliefs and depressive symptoms. Second, we employed three econometric approaches with relatively weaker, (i.e., more likely to hold) identifying assumptions in order to (a) mitigate issues of selection bias and (b) conduct within-study replications or robustness checks to assess the extent to which our results were consistent across different methods and models (Duncan et al., 2014). Finally, we employed, for the first time in this body of research, falsification tests aimed at assessing the internal validity of the findings.

Besides methodological contributions, this study made two major conceptual contributions. First, this study focused on the effects of physical punishment on cognition, a relatively understudied outcome (Gershoff & Grogan-Kaylor, 2016). Second, this research used a sample from Colombia, a country where children are exposed to high levels of violence, including physical punishment. Indeed, three out of five 3- to- 4-year-olds and one out of 10 children in their first year of life are physically punished in Colombia (Cuartas, 2018; Cuartas, McCoy, et al., 2019) and Colombia is one of the few Latin American countries without a legal ban on physical punishment (Global Initiative to End Corporal Punishment of Children, 2019). A better understanding of the potential effects of physical punishment in the context of Colombia will be useful to demonstrate the potential replicability and generalizability of findings from previous studies conducted in the U.S. and to inform potential policy and practice interventions aimed at protecting children from developmental risk factors.

Method

Participants

Data for the present study came from a cluster-randomized controlled trial (RCT) conducted in Colombia (Attanasio et al., 2014). The RCT was aimed at evaluating the impact of the delivery of psychosocial stimulation and micronutrient supplementation to a sample of children younger than three years and their mothers taken from the poorest quintile of households in the country. Prior analysis of the RCT found no effect of the interventions on parents' use of physical punishment and positive effects on children's cognition (Attanasio et al., 2014). Preintervention data (Time 1, or T1) were gathered between February and June 2010, when children were on average 17.78 months old (i.e., 1.5 years; range = 9–26 months). Postintervention data (Time 2, or T2) were gathered between September and December 2011, when children were on average 36.56 months old (i.e., 3 years; range: 27–46 months). The data include mothers' reports of child, family, and household characteristics, and selected direct measures of children's physiology, anatomy, and development. We combined these data with the Panel Municipal del CEDE (Acevedo & Bornacelly, 2014), which is a longitudinal dataset compiled by the Faculty of Economics at Universidad de los Andes with information about the demographic, social, economic, and political characteristics of the municipality (i.e., smallest administrative unit in Colombia). We did not seek Harvard University's Institutional Review Board (IRB) approval for the current study because it employed secondary, de-identified, publicly available data.

Our analytical sample was comprised of 1,167 child-mother pairs without missing information on the key study variables, living in 95 municipalities (see Supplemental Figure S1 for map of included municipalities). These 1,167 child-mother pairs constitute 83.66% of the 1,395 pairs included in in the full sample at T1. As such, we had 16.4% missing cases, which is roughly equivalent between children exposed to physical punishment (16.9%) and not exposed (15.4%). Moreover, the analytic sample was not statistically significantly different, on average, from the excluded sample in terms of key variables, suggesting that the missingness was not systematic (see Supplemental Table S1 for details). As shown in Table 1, 51% of children in the analytic sample were boys and only 67% lived with their fathers. Mothers averaged 26.2 years old (range = 14–47 years) and 35 had a primary education or less.

Measures

Cognitive skills.—The Bayley Scales of Infant and Toddler Development, Third Edition (Bayley-III), were used to directly assess children's cognitive, receptive language, expressive language, fine motor, and gross motor development. The Bayley-III is administered individually by trained examiners who present a set of materials (e.g., toys, memory cards) and tasks to the child and scores the child's responses. For example, the examiner may present the child with a simple puzzle and ask him or her to complete the puzzle, scoring based on how many pieces are correctly placed. Before administering the test, examiners ask parents not to interact with the child during the assessment to avoid distractions or unwarranted interferences that could bias the results. The test takes around 30–90 min. There are 72 items in the cognitive subscale, 49 items in the receptive language

subscale, 48 items in the expressive language subscale, 66 in the fine motor subscale, and 72 items in the gross motor subscale, with start and stop rules based on age and task completion. (For more details on test administration and scoring, see Albers & Grieve, 2007; Bayley, 2006a, 2006b).

The Bayley-III was translated into Spanish following a translation and back translation process to ensure accuracy and was implemented by trained testers with degrees in psychology in local community centers in the presence of mothers (Attanasio et al., 2014). The Bayley-III exhibited adequate psychometric properties in Colombia, with interrater reliability above 0.90 in this sample (Attanasio et al., 2014) and test–retest reliability of 0.95 to 0.98 in another sample from Bogotá, Colombia's capital city (Rubio-Codina, Attanasio, Meghir, Varela, & Grantham-McGregor, 2015). Moreover, a review of measurement tools for child development in Colombia and Latin America concluded that the Bayley-III exhibited the best sensitivity, specificity, predictive validity and reliability to measure child development among a set of measurement tools used in the region (Jurado Castro & Rebolledo-Cobos, 2017). Because the reference population for the composite scores is a sample of U.S. children which may not be appropriate in Colombia, we followed Attanasio et al. (2014) and used raw scores throughout our analyses (controlling for age). Due to the fact that the Bayley-III is a widely used direct assessment with adequate psychometric properties in cross-cultural contexts, our reliance on the Bayley-III eliminates concerns regarding shared method variance (e.g., Ferguson, 2013). This study focuses on the Bayley-III cognitive raw score as the main outcome of interest and uses the remainder of the subscales as covariates (see Table 1 for descriptive statistics).

Physical punishment.—Mothers responded using "yes" or "no" to the following question from a survey: "During the last seven days, when the child misbehaved or did things that bothered you, did you have to hit him or her?" (Original item in Spanish: "Durante los últimos 7 días contados hasta ayer, cuando [niño/niña seleccionado(a)] se portó mal o hizo cosas que no estaban bien o que a usted le molestaron, ¿tuvo que pegarle?"). In Colombia, "tuvo que pegarle" (literal translation: "had to hit him/her") normally refers to spanking or smacking the child (ICBF, 2017). According to their reports, 41% of sampled mothers employed physical punishment in the seven days before the survey took place, which is consistent with the prevalence of low-income mothers' stated regular use of physical punishment found in previous studies that have used representative samples for the same geographic regions (Cuartas, 2018).

Covariates.—Table 1 summarizes descriptive statistics for all variables included in the study. We included as covariates an extensive set of characteristics of the children, their mothers, the home environment, and their municipality, all measured at T1. We included *children's age in months*, *sex*, *hemoglobin concentration* (g/dl) as a marker for anemia, *weight* in kg (accurate to 0.1 kg), and *height* in cm (accurate to 1 mm). To control for children's overall developmental level, we included Bayley-III raw scores for *cognitive skills* (to create a lagged dependent variable), *receptive language, expressive language, fine motor skills,* and *gross motor skills*. As markers of children's temperament, we included three ratings done by trained psychologists at the end of the Bayley-III assessment,

following the 5-point rating scale developed by Wolke, Skuse, and Mathisen (1990). These observers rated children's *activity level*, which characterized how physically active the child was during the testing, their *emotional tone*, which referred to their observed affective state, and their level of *cooperativeness*, which reflected how much the child cooperated with the examiner during the Bayley-III administration. These measures have been found to be accurate measures of children's behavior and temperament when contrasted with parental reports (Hamadani et al., 2010).

Maternal characteristics were *age in years*, *highest level of education* (primary vs. secondary or higher), and whether they were *currently attending school*. We included maternal *depressive symptoms*, which were measured using the Spanish version of the short version of the Center for Epidemiologic Studies Depression Scale (CES-D-10), which is a 10-item scale to identify symptoms associated with depression in the past week (Radloff, 1977). Sample items included, "I felt lonely" and "I felt hopeful about the future" (reversed). The CES-D-10 had adequate internal consistency with a Cronbach's alpha (α) of .79 in our sample, and overall (sum) scores had a hypothetical range from zero to 60. We also included an indicator of mothers' *level of disability* which was indexed by their self-reported difficulty in walking, doing physical activity, communicating with others, dressing by themselves, and performing cognitive tasks (observed range = 0 to (1). Additionally, we included nine yes (1) or no (0) indicators for different *maternal beliefs regarding child development*: (a) children's intelligence changes very little after birth, (b), children who know more words learn to read earlier, (c) children who play a lot with their mothers and other children have higher achievement at school, (d) no matter what a mother does, each child starts to talk according to his or her own nature, (e) children with higher achievement at school earn more money in their adulthood, (f) it is good that children spend a lot of time playing alone, (g) it is better to wait until children aged 12–24 months old understand what someone says to them before telling them stories and tales, (h) praising or applauding a child too much makes him or her overly confident, and (i) it is important that a busy parent spends a lot of time playing with his or her child.

To characterize the home environment, we included whether the child's *father lived at home* and whether a member of the *household or family had recently died*. An index of *family wealth* was computed from mothers' reports of dwelling characteristics (e.g., wall and floor materials) and assets (e.g., TV, fridge, washing machine). Following the algorithm presented in Filmer and Pritchett (2001), we used principal component analysis to compute a wealth index to characterize levels of multidimensional poverty. This index ranged from −10.87 to 3.19. To describe the learning environment of the home, we followed Hamadani and colleagues' (2010) procedures for doing so in underdeveloped countries by computing an index for availability of *stimulating materials* (α = 60), including the number of toys, blocks, and books available at home (possible range = 0 to 9). Moreover, we computed an index of caregiver *cognitive stimulation*, similar to the one used by Bornstein and Putnick (2012), by counting the number of activities that mothers reported they engaged in with the child in the three days preceding the survey. The index included caregiver-child activities such as reading books, telling stories, singing songs, doing outdoor activities, playing with toys, drawing, and naming or counting objects (possible range = 0 to 7;α = .66).

Finally, to characterize the municipality in which families lived, we included the *total population*, *number of resources* (e.g., schools, hospital, child care centers), *number of shocks* (e.g., drought, flood, plague), the *multidimensional poverty index* (Alkire & Foster, 2011) using data from the most recent national census (2005), and the *unsatisfied basic needs index* (Feres & Mancero, 2001) for 2011. We also included the following municipality-level information for the five years preceding the survey, taken from the Panel Municipal del CEDE (Acevedo & Bornacelly, 2014): *average Saber 11 test score* (a national standardized test to measure academic achievement), *presence of FARC guerrillas* (1 = yes; 0 = no), and the *homicide rate*, *theft rate*, and *terrorism rate,* each calculated per 100,000 inhabitants as reported by the national government.

Analysis and Identifying Assumptions

Equation 1 presents a basic model to analyze the association between physical punishment and children's cognitive development. In this model, *cognitive* is the raw score on the Bayley-III cognition subscale for child *i* at T2, *physical* is the predictor of interest and indicates whether child *i* was exposed to physical punishment in T1, *covariates* is a set of V covariates at T1 to reduce the risk of confounding, and ε represents residual variation.

$$\text{cognitive}_i^{T2} = \alpha_0 + \beta \text{physical}_i^{T1} + \sum_{v=1}^{V} \gamma_v \text{covariates}_i^{T1} + \varepsilon_i \quad (1)$$

In this model, β represents the association between physical punishment at T1 and children's cognitive development raw score at T2. This regression coefficient will represent a causal effect under the unlikely scenario that the groups of children exposed and not exposed to physical punishment are, on average, equal at T1 (as if random assignment had taken place) or if *covariates* include all potential confounders and, consequently, the residual term is uncorrelated with the outcome variable or main predictor. As shown in Table S2, there are several observed differences at T1 between children exposed and not exposed to physical punishment—including preexisting differences in cognitive skills—which makes this assumption unlikely to hold. Figure S2 shows the distribution of scores by physical punishment group at both T1 and T2; as seen in T1, there were initial differences in cognitive skills at baseline, such that children in the physical punishment group had higher cognitive scores. We followed three different strategies aimed at controlling for these preexisting differences.

For our first set of analyses we used lagged dependent variable estimates (Angrist & Pischke, 2011), including child *i's* cognitive development score at T1 as an additional control variable (see Equation 2). Indeed, prior cognitive scores may be the best predictor of later cognitive scores and may capture several individual and ecological factors (e.g., biological characteristics, caregivers' warmth and stimulation) that predict children's overall cognitive developmental trajectory, largely reducing the risk of confounding. This specification will produce an unbiased estimate of the average treatment on the treated (ATT) under the conditional independence assumption, which in this case would be that the

variation left unexplained after accounting for cognitive raw scores at T1 and other control variables (i.e., ε_i) does not correlate simultaneously with the outcome variable and the main predictor.

$$\text{cognitive}_i^{T2} = \alpha_0 + \beta\text{physical}_i^{T1} + \varphi_1\text{cognitive}_i^{T1} + \sum_{v=1}^{V} \gamma_v\text{covariates}_i^{T1} + \varepsilon_i \quad (2)$$

Our second analytical approach is a difference-in-differences-like approach (DD) following the model presented in Equation 3. The traditional DD is a statistical method applied to longitudinal data that aims to control for unobserved time-invariant differences between two groups often produced by a quasi-experiment or natural experiment, being similar to fixed-effects models (Blundell & Dias, 2009; Foster, 2010). Because we do not have a natural experiment to capitalize upon in our study, we compare changes in cognitive development over time (first difference) across the physically punished and not physically punished groups (2nd difference), thereby making it akin to a fixed effects model. We will refer to this as a "DD-like" method. This DD-like method involves the subtraction of the average change in the outcome variable between T1 and T2 from the first group (in this case, children exposed to physical punishment) from the average change in the other group (children not exposed to physical punishment). In our model, which follows the DD procedures, *Time* is an indicator for the wave of data collection that equals one for T2. In addition, α_1 represents the average difference in cognitive development between children exposed and not exposed to physical punishment at T1 and α_2 represents the average change in cognitive development scores for the overall sample from T1 to T2.β, the coefficient for the interaction term between *physical* and *time*, captures the "double-difference" and represents the average change in cognitive development from T1 to T2 for children exposed to physical punishment, relative to those who were not exposed.

$$\text{cognitive}_i = \alpha_0 + \alpha_1\text{physical}_i^{T1} + \alpha_2\text{time} + \beta\left(\text{physical}_i^{T1} \times \text{time}\right) + \sum_{v=1}^{V} \gamma_v\text{covariates}_i^{T1} + \varepsilon_i \quad (3)$$

DD will allow us to obtain an unbiased estimate of the ATT under the assumption of parallel trends (Blundell & Dias, 2009), or the assumption that the natural trend of cognitive development for children who have been physically punished would, had physical punishment not taken place, have been the same as that for children not exposed to physical punishment. We added a set of covariates at T1 to improve model efficiency and to evaluate the robustness of the estimates. (Adding covariates at T2 may bias the estimates, as these may be affected by physical punishment at T1.) As discussed by Larzelere et al. (2018) and Angrist and Pischke (2011), robust consistency in results across models with lagged outcome variables versus DD offers additional evidence on the credibility of an unbiased causal effect. Yet, taking into account the dynamic, transactional set of systems in which

parenting takes place and human development unfolds (Bronfenbrenner & Morris, 2007; McCoy, 2013; Sameroff, 2010), the risk of omitting relevant time-varying variables correlated both with physical punishment and children's development is considerable, potentially biasing the DD estimates. This risk is further compounded by the fact that it is parents that self-select into employing physical punishment; there is no exogenous event that predicts physical punishment usage.

Our third and last methodological approach is a difference-in-differences-like approach with matching (DDM), which we used in an attempt to mitigate the possibility of the threat to validity posed by omitted time-varying variables. Similar to DD, DDM subtracts the average change in the outcome variable and performs a second difference between a matched (i.e., similar in observed characteristics) sample of children exposed and not exposed to physical punishment to estimate the ATT effect. DDM, consequently, controls for time-invariant characteristics (as does DD alone) while restricting the comparison to children with the most similar observed characteristics (as does PSM alone).

We followed the steps suggested by Bernal and Peña (2011) and Caliendo and Kopeinig (2008) to perform the PSM. First, we used variables that, according to theoretical models and empirical evidence, may potentially confound the relation between physical punishment and children's cognitive development to estimate the propensity score (i.e., probability of being exposed to physical punishment). Considering evidence from Colombia about predictors of physical punishment (e.g., Cuartas, Grogan-Kaylor, Ma, & Castillo, 2019), we used the following variables at T1 to conduct the matching: child's age, sex, expressive language raw score, hemoglobin levels, weight, and the three observed measures of temperament; mother's age, depressive symptoms, index of stimulation, availability of stimulating materials, wealth index, indicator for father lives at home, and municipality unmet basic needs, index for resources and shocks, and homicides and terrorism rates.

Subsequently, we restricted the sample to the common support, or the observations exhibiting positive probabilities of being both exposed and not exposed to physical punishment, by deleting all observations whose propensity score was smaller than the minimum or larger than the maximum in the other group (Caliendo & Kopeinig, 2008). Then, we used four different algorithms to create matched samples: (a) nearest neighbor matching, (NN), which compares each individual in the physical punishment group with the most similar individual in the control group according to observed characteristics summarized in the propensity score, (2) four nearest neighbors (4-NN), with compares each individual in the physical punishment group with the four most similar individuals in the control group, (3) radius, which imposes a tolerance level for the maximum propensity score distance defining the comparison groups (in our case a caliper of 0.05), and (4) kernel matching, which assigns weights based on the distance between the propensity score of each observation in each group (i.e., exposed and not exposed to physical punishment) to make a weighted comparison between all observations. As an additional robustness check, we employed entropy matching, a different matching method that is not based on propensity scores but on reweighting schemes (Hainmueller, 2012). This approach allows us to adjust the distribution for both the mean and variance of each variable in the study.

We assessed the quality of each matching technique by examining the differences in the means of all observed characteristics (including those not used to predict the propensity score) across children exposed and not exposed to physical punishment to analyze if balance was achieved. Finally, we estimated standard errors using bootstrapping (i.e., reestimating the effect *N* times) with 100 repetitions to obtain more accurate asymptotic inferences, given that PSM tends to understate standard errors (Caliendo & Kopeinig, 2008).

In sum, combining matching and a DD-like approach (i.e., DDM) has several strengths. First, matching serves as a diagnostic tool (Foster, 2010), allowing us to assess whether the matched sample is, on average, comparable in observed characteristics at T1. Second, matching, in particular PSM, serves as a restricting tool, excluding from the analyses children who were not exposed to physical punishment who are not comparable in observed characteristics to children who were. Finally, and most importantly, combining DD and matching allows us to relax, to a certain extent, the identifying assumption of each method. In particular, DD controls for unobserved characteristics that do not vary with time, relaxing matching's unlikely conditional independence assumption (i.e., that omitted variables are not correlated with physical punishment or cognitive development), whereas matching with a rich set of T1 confounders produces a more comparable comparison group for the punishment-exposed children. Ultimately, our DDM approach will produce unbiased estimates of the effect of physical punishment on children's cognitive development under the assumption that unexplained time-variant heterogeneity is not correlated simultaneously with cognitive development and physical punishment simultaneously (which is a considerably weaker identifying assumption in comparison to matching or DD, yet still a potential concern).

Finally, we conducted a falsification test to assess the internal validity of our estimates. In particular, given that a natural condition of a causal effect is that the cause must precede the effect, replicating our analyses in reversed time should not replicate the observed effects (Larzelere et al., 2018). We employed lagged and DDM models to examine whether physical punishment at T2 predicted cognitive development at T1. In this case, finding a statistically significant association between earlier cognitive development and later physical punishment would be a sign of a spurious correlation or endogeneity bias (probably omitted variable bias), weakening our overall argument.

Lagged and DD models were estimated using clustered-standard errors at the municipality-level to account for the sampling design (Attanasio et al., 2014), whereas bootstrapped standard errors were estimated for PSM models (Caliendo & Kopeinig, 2008). All analyses were conducted in Stata/MP 15.1.

Results

Table 2 summarizes the results for three versions of the lagged dependent variable model presented in Equation 2, which regresses T2 cognitive scores on T1 physical punishment, T1 cognitive scores, and a set of T1 covariates that were included in a stepwise fashion (from Column 2 to 4 in Table 2) to assess the robustness of the estimates. (Table S3 presents coefficients for all covariates.) Before describing these results, Column 1 presents the

bivariate correlation between physical punishment and cognitive scores ($\beta = -0.22$, $SE = 0.06$, $p < .001$). The most basic model (Column 2) controls only for children's T1 cognitive development and their age and sex and finds that children who were physically punished exhibited cognitive scores at T2 that were 0.08 SD ($SE = 0.04$, $p < .10$) lower than their peers who were not exposed to physical punishment. Including additional control variables (Columns 3 and 4) increases the precision of the estimates, $\beta = -0.09$, $SE = 0.04$, $p < .05$, and shows that the estimated association between physical punishment and children's cognitive development is robust to the inclusion of potential confounders, including cognition at T1.

Table 3 presents results for the DD-like model presented in Equation 3. (Table S4 presents coefficients for all covariates). As expected, children developed more cognitive skills over the approximately 18 months between T1 and T2, even with all covariates included ($\alpha_2 = 1.79$, $SE = 0.02$, $p < .01$). An unexpected finding was that physical punishment at T1 was contemporaneously associated with more cognitive skills at T1, even when our extensive set of covariates was included ($\alpha_1 = 0.11$, $SE = 0.02$, $p < .01$). However, when change among those experiencing physical punishment was compared with change among those who did not (the interaction term in Table 3), physical punishment at T1 predicted significantly lower scores on Bayley-III cognition at T2, even after accounting for unobserved time-invariant variables (i.e., using DD) and our set of observed covariates ($\beta = -0.19$, $SE = 0.03$, $p < .01$).

Figure 1, Panel A, presents the common support once the propensity score was estimated at T1 using a logit model. The density graph shows that most observations fall within the area of common support (denoted by the dotted lines), suggesting that PSM estimates are feasible. Moreover, the overlapping distributions in panels B to F show that different PSM algorithms and entropy matching produce balance in the propensity scores across children who were exposed to physical punishment and those who were not. Table S5 presents further tests for the quality of the matching techniques, revealing that balance was achieved in all observed individual, household, and municipality characteristics at T1 between children exposed and not exposed to physical punishment using different PSM algorithms and entropy matching.

Table 4 presents the observed coefficients for the association between physical punishment and cognitive development using DDM with the five different matching approaches. Across each specification, T1 physical punishment predicts slower gains in children's cognitive development, even after accounting for unobserved time-invariant characteristics and restricting the comparison to exposed and unexposed children who were similar according to a comprehensive set of observed characteristics. More importantly, the magnitude and statistical significance of this association is robust to different matching algorithms, with an estimated coefficient that falls between $\beta = -0.12$, $SE = 0.06$, $p < .01$, for entropy matching to $\beta = -0.21$, $SE = 0.08$, $p < .05$ for PSM using the nearest neighbor algorithm.

Summary of Results

Figure 2 summarizes the main results from the study, taken from three empirical strategies and several within-strategy robustness checks. The bars represent 95% confidence intervals around the estimates for each method; bars that do not cross zero in the figure are

statistically different from zero. Overall, exposure to physical punishment at ages 9–26 months predicts lower growth in cognitive ability about 2 years afterward, with an estimated difference of between −0.08 to −0.21 *SD*, depending on the empirical strategy employed.

Falsification Test

Finally, in Figure 3 we provide additional evidence about the internal validity of our estimates from a falsification test similar to one reported in Larzelere et al. (2018), namely, we tested a model with the false assumption that T2 physical punishment would "predict" back in time to T1 cognitive skills. The figure shows that physical punishment at T2 does not predict (i.e., all bars cross zero) lower cognitive skills at T1 in any of the six model specifications we used above. (See Table S6 for details). This finding provides additional support to the claim that physical punishment predicts slower growth in cognitive skills over time.

Discussion

The aim of this study was to provide evidence on the plausibility of an effect of physical punishment on cognitive development during early childhood. The findings suggest that Colombian children who were exposed to physical punishment at ages 9–26 months demonstrated cognitive skills at 27 – 46 months that were 0.08–0.21 *SD* lower than their peers who were not exposed, even after accounting for previous levels of cognitive development, a comprehensive set of child, mother, home environment, and municipality characteristics, and time-invariant heterogeneity. Falsification tests (i.e., predicting previous cognitive development with future physical punishment) provided further evidence for the validity of the estimates.

Substantively, findings from this study contribute to and are consistent with the body of literature linking physical punishment to detriments in child cognition. Across our seven different model specifications, our effect sizes ranged from −0.08 to −0.21. This range includes the mean effect size for the association between physical punishment and cognitive skills from the meta-analysis of 8 studies by Gershoff and Grogan-Kaylor (2016), $d = -0.17$ (95% CI [−.01, −.32]), as well as the mean effect size for the same association from the meta-analysis of 4 studies by Ferguson (2013), $d = -0.11$ (95% CI [−.05, −17]). This consistency in the magnitude of these effect sizes is remarkable given that our use of both DD and matching is more rigorous than the methods of Gershoff and Grogan-Kaylor (2016), who relied on bivariate and often cross-sectional associations, and of Ferguson (2013), who only controlled for initial levels of cognitive skills. Our findings thus lend additional confidence in the findings of these previous meta-analyses by demonstrating that the association between physical punishment and slower growth in cognitive skills is robust to concerted efforts to address omitted variable bias.

We were not able to directly test our two hypothesized mediational pathways to explain links between physical punishment and cognitive skills. First, we did not have measures of children's autonomic nervous system responses to being physically punished, which we would need to test our stress-based hypothesis. This will be hard to achieve in future research, as parents use physical punishment relatively rarely (typically once a week or less)

and so a naturalistic assessment of their physical stress responses would be difficult. Continued brain scan studies and creative experiments like that by Bugental and colleagues (2003) may be our best bet for truly understanding how physical punishment "gets under the skin" of children. Second, we did not have measures of whether the child behaviors parents chose to respond to with physical punishment were behaviors that foster cognitive development. Intensive, in-home observations of the child behaviors that trigger physical punishment would be needed—although, again, such a strategy may not be feasible given that physical punishment does not occur every day or even every week. Third, we did not have three waves of data that we would have needed to temporally separate the predictor, mediator, and outcome in order to establish our hypothesized causal direction. In short, our findings are consistent with theory and past research and confirmed that physical punishment predicts slower cognitive growth; they just cannot tell us *why*.

Given the current results and prior evidence, we were initially surprised by the finding in the DD models that physical punishment at T1 was contemporaneously related to higher cognitive skills at T1 (see Table 3). Importantly, however, we do not believe this bivariate relation to be causal. Instead, because physical punishment and child cognition were measured at the same time, it is possible that children with higher cognitive skills are eliciting more physical punishment from their parents (perhaps because of a need for stimulation), or that a third, unmeasured variable predicts both physical punishment and cognitive skills in toddlers (as both were measured at the same time). We also compared our findings with those of Berlin et al. (2009), who used the same measure of cognitive skills as in the current study (the Bayley-III) with the same-aged children and with a similarly large and disadvantaged sample in the United States. Berlin and colleagues did not find spanking to be correlated within time with cognitive skills when children were aged 2 or 3, in contrast to our finding of an association at 18 months. They did not find significant links between spanking at age 1 and cognitive skills at age 2, nor between age 2 spanking and age 3 cognitive skills. They also did not find any evidence of a child evocative effect: age 2 cognitive skills did not predict age 3 spanking (Berlin et al., 2009). However, Berlin and colleagues did find a longitudinal link between age 1 spanking and significantly lower cognitive skills at age 3 ($\beta = -.06$, $p < .05$), just as we did between spanking at (on average) age 1.5 and lower cognitive skills at age 3. The fact that our study replicated this earlier study in finding links between spanking and cognitive skills 1.5 to 2 years later suggests either that the effects of physical punishment on children's cognitive development accumulate over time or that physical punishment has a greater impact on skills that develop in the preschool years rather than those that develop in infancy and toddlerhood. Future research with more closely spaced assessments would be needed to determine when these associations begin to manifest.

Methodologically, our analyses build upon recent longitudinal studies using fixed-effects models (Ma et al., 2018) and PSM (Gershoff et al., 2018) to provide more internally valid estimates of the potentially harmful effects of physical punishment on child development. Importantly, we directly respond to previous work in the developmental literature (e.g., Duncan et al., 2014; Larzelere et al., 2018; Miller et al., 2016) by conducting within-study and within-model replications to assess the robustness of the estimates and by testing the validity of the identification strategy using a falsification test. Collectively, these approaches

offer promise not only for improving the field's understanding of the effects of physical punishment on children's outcomes, but also for serving as a potential template for future developmental studies probing causal questions using observational data sets.

Implications for Policy and Practice

More than 60% of young children around the world are physically punished by their parents (Cuartas, McCoy, et al., 2019), despite the large body of mostly correlational evidence linking physical punishment to harm to children (e.g., Ferguson, 2013; Gershoff, Goodman, et al., 2018; Gershoff & Grogan-Kaylor, 2016; Ma et al., 2018). Our study joins several recent studies (e.g., Gershoff et al., 2018; Ma et al., 2018; Okuzono et al., 2017) in using advanced statistical methods that improve the field's ability to conclude that physical punishment is in fact the cause of detrimental child outcomes. Given that experiments assigning children to be hit by their parents are unethical, studies such as ours that mitigate concerns about omitted variables using methods with more plausible identifying assumptions and that conduct falsification tests further strengthen causal arguments.

The strength and consistency of the finding that physical punishment harms children imply that approaches to reduce caregivers' use of physical punishment and to increase alternative, non-violent methods to modify children's behavior and support their development are needed in Colombia and elsewhere (see Durrant, 2016; Sege, Siegel, & Council on Child Abuse and Neglect and the Committee on Psychosocial Aspects of Child and Family Health, 2018). More than 50 countries worldwide have implemented legal bans on physical punishment as a means of educating their citizens about the potential harms of physical punishment and thereby reducing its use over time (Global Initiative to End Corporal Punishment of Children, 2019). Colombia, a nation where children's exposure to domestic and community violence is dramatically high (Save the Children, 2017), is not one of these countries. Moving forward, countries like Colombia may consider such bans as one means of discouraging physical punishment and encouraging more positive disciplinary techniques (Lansford et al., 2017).

Importantly, any legal prohibitions should not be punitive, as sanctions against caregivers may serve as an additional contextual source of stress or fear for children and families. Instead, evidence suggests that a strengths-based approaches whereby additional resources are provided to assist parents and families in supporting their young children's healthy development may serve as more effective means of behavior modification (Gershoff, Lee, & Durrant, 2017). More research is needed to understand culturally specific positive disciplinary methods that could be encouraged in place of physical punishment.

Limitations and Future Directions

The present study has important limitations that must be noted. First, although our statistical methods improve the inference of the relations between physical punishment and children's cognitive development, threats to validity remain. For example, for our DD-like approach we did not exploit an exogenous event (e.g., a natural disaster or a policy-shift) that led parents to employ physical punishment, a methodological strategy that is often utilized when DD is used in order to control for the fact that it is parents who self-selected into employing

physical punishment. Similarly, in the DDM models, unobserved time-varying confounders not included to match the sample are likely to bias the estimates. However, the robustness of our results, and evidence from the falsification tests, mitigate these concerns and support the argument that the estimates are internally valid. Second, although the measure of physical punishment used in this study is common in the field (see Gershoff & Grogan-Kaylor, 2016) and was intended to capture caregivers' behaviors in the seven days before the survey, it is possible that children were physically punished before that period of time or that parents underreported their actual use of physical punishment due to social desirability. Either issue would produce an underestimation of children's actual exposure to physical punishment and the true effects of physical punishment exposure. These estimates are thus likely lower bounds or conservative estimates for the potential effect of physical punishment on cognitive development. Similarly, the measure of physical punishment does not allow us to assess differential effects according to frequency or severity of exposure.

Third, the study focused exclusively on physical punishment inflicted by mothers, who are mostly the primary caregivers of children in Colombia (Ministerio de Salud & Profamilia, 2016), and does not consider other caregivers such as fathers. Future studies should examine the independent and interactive effects of corporal punishment inflicted by different caregivers on the development of young children. Additionally, this study only provided evidence for the short-term (less than 2-year) effects of physical punishment on cognitive development. Future research should explore long-term consequences of punishment on cognitive, behavioral, and other developmental outcomes, as well as the more acute effects of a single incident. Finally, more studies are needed in order to understand how generalizable the findings from this study are to other settings and populations both inside and outside of Colombia.

Conclusion

Establishing causality is essential in developmental psychology in general and in research about physical punishment in particular. Experimental studies are often not feasible when studying how aspects of children's environments affect their development, and thus researchers must rely on alternative methods with stronger identifying assumptions that make causal claims less convincing. Developmentalists should continue to invest in finding opportunities to leverage natural variability that allow us to study meaningful research questions with better identification, while relying on strong theory, weaker identifying assumptions, and tests of the robustness and replicability of results to alternative methods, models, and populations. Building on the results from this study as well as recent writings about causality in developmental science (e.g., Duncan et al., 2014; Foster, 2010; Miller et al., 2016), we encourage the continued use of quasi-experimental approaches, such as fixed-effects designs, instrumental variables, regression discontinuity-designs, and DDM, as well as the use of multiple sensitivity, robustness, and falsification tests to further strengthen causal inference in research on the consequences of physical punishment for children's development.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

Acevedo K, & Bornacelly I. (2014). Panel municipal del CEDE [Municipal Panel Data CEDE] (Documentos CEDE No. 26). Bogotá, Colombia: Universidad de los Andes.

Afifi TO, Ford D, Gershoff ET, Merrick M, Grogan-Kaylor A, Ports KA, . . . Peters Bennett R. (2017). Spanking and adult mental health impairment: The case for the designation of spanking as an adverse childhood experience. Child Abuse & Neglect, 71, 24–31. 10.1016/j.chiabu.2017.01.014 [PubMed: 28126359]

Alampay LP, Godwin J, Lansford JE, Bombi AS, Bornstein MH, Chang L, . . . Bacchini D. (2017). Severity and justness do not moderate the relation between corporal punishment and negative child outcomes: A multicultural and longitudinal study. International Journal of Behavioral Development, 41, 491–502. 10.1177/0165025417697852 [PubMed: 28729751]

Albers CA, & Grieve AJ (2007). Test review: Bayley N (2006). Bayley Scales of Infant and Toddler Development–Third Edition. San Antonio, TX: Harcourt Assessment. 10.1177/0734282906297199

Alkire S, & Foster J. (2011). Counting and multidimensional poverty measurement. Journal of Public Economics, 95, 476–487. 10.1016/j.jpubeco.2010.11.006

Angrist J, & Pischke J-S (2011). Mostly harmless econometrics: An empiricist's companion. Princeton, NJ: Princeton University Press.

Attanasio OP, Fernández C, Fitzsimons EOA, Grantham-McGregor SM, Meghir C, & Rubio-Codina M. (2014). Using the infrastructure of a conditional cash transfer program to deliver a scalable integrated early child development program in Colombia: Cluster randomized controlled trial. British Medical Journal, 349, g5785. 10.1136/bmj.g5785 [PubMed: 25266222]

Bayley N. (2006a). Bayley Scales of Infant and Toddler Development–Third Edition: Administration manual. San Antonio, TX: Harcourt Assessment.

Bayley N. (2006b). Bayley Scales of Infant and Toddler Development–Third Edition: Technical manual. San Antonio, TX: Harcourt Assessment

Berlin LJ, Ispa JM, Fine MA, Malone PS, 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. 10.1111/j.1467-8624.2009.01341.x [PubMed: 19765008]

Bernal R, & Peña X. (2011). Guía práctica para la evaluación de impacto [Practical guide for impact evaluation]. Bogotá, Colombia: Ediciones Uniandes.

Berry D, & Willoughby MT (2017). On the practical interpretability of cross-lagged panel models: Rethinking a developmental workhorse. Child Development, 88, 1186–1206. 10.1111/cdev.12660 [PubMed: 27878996]

Blundell R, & Dias MC (2009). Alternative approaches to evaluation in empirical microeconomics. The Journal of Human Resources, 44, 565–640. 10.3368/jhr.44.3.565

Bornstein MH, & Putnick DL (2012). Cognitive and socioemotional caregiving in developing countries. Child Development, 83, 46–61. 10.1111/j.1467-8624.2011.01673.x [PubMed: 22277006]

Bronfenbrenner U, & Morris P. (2007). The bioecological model of human development. In Lerner RM & Damon W (Eds.), Theoretical models of human development (pp. 793–828). Hoboken, NJ: Wiley.

Bugental DB, Martorell GA, & Barraza V. (2003). The hormonal costs of subtle forms of infant maltreatment. Hormones and Behavior, 43, 237–244. 10.1016/S0018-506X(02)00008-9 [PubMed: 12614655]

Caliendo M, & Kopeinig S. (2008). Some practical guidance for the implementation of propensity score matching. Journal of Economic Surveys, 22, 31–72. 10.1111/j.1467-6419.2007.00527.x

Coulton CJ, Crampton DS, Irwin M, Spilsbury JC, & Korbin JE (2007). How neighborhoods influence child maltreatment: A review of the literature and alternative pathways. Child Abuse & Neglect, 31, 1117–1142. 10.1016/j.chiabu.2007.03.023 [PubMed: 18023868]

Cuartas J. (2018). Physical punishment against the early childhood in Colombia: National and regional prevalence, sociodemographic gaps, and ten-year trends. Children and Youth Services Review, 93, 428–440. 10.1016/j.childyouth.2018.08.024

Cuartas J, Grogan-Kaylor A, Ma J, & Castillo B. (2019). Civil conflict, domestic violence, and poverty as predictors of corporal punishment in Colombia. Child Abuse & Neglect, 90, 108–119. 10.1016/j.chiabu.2019.02.003 [PubMed: 30772750]

Cuartas J, McCoy DC, Rey-Guerra C, Britto PR, Beatriz E, & Salhi C. (2019). Early childhood exposure to non-violent discipline and physical and psychological aggression in low- and middle-income countries: National, regional, and global prevalence estimates. Child Abuse & Neglect, 92, 93–105. 10.1016/j.chiabu.2019.03.021 [PubMed: 30939376]

Danese A, & McEwen BS (2012). Adverse childhood experiences, allostasis, allostatic load, and age-related disease. Physiology & Behavior, 106, 29–39. 10.1016/j.physbeh.2011.08.019 [PubMed: 21888923]

Duncan GJ, Engel M, Claessens A, & Dowsett CJ (2014). Replication and robustness in developmental research. Developmental Psychology, 50, 2417–2425. 10.1037/a0037996 [PubMed: 25243330]

Duncan GJ, Magnuson KA, & Ludwig J. (2004). The endogeneity problem in developmental Studies. Research in Human Development, 1, 59–80. 10.1207/s15427617rhd0101&2_5

Durrant J. (2016). Positive discipline in everyday parenting. Save the Children Sweden Resource Centre. Retrieved from http://www.positivedisciplineeveryday.com

Feres JC, & Mancero X. (2001). El método de las necesidades básicas insatisfechas (NBI) y sus aplicaciones en América Latina [The unsatisfied basic needs approach and its applications in Latin America]. Santiago de Chile, Chile: Estudios estadísticos. CEPAL.

Ferguson CJ (2013). Spanking, corporal punishment and negative long-term outcomes: A meta-analytic review of longitudinal studies. Clinical Psychology Review, 33, 196–208. 10.1016/j.cpr.2012.11.002 [PubMed: 23274727]

Filmer D., & Pritchett LH. (2001). Estimating wealth effects without expenditure data—Or tears: An application to educational enrollments in states of India. Demography, 38, 115–132. 10.2307/3088292 [PubMed: 11227840]

Fitzsimons E, Rubio-Codina M, Andrew A, & Attanasio O. (2010). Colombia medium-term effects of home-based early childhood development intervention impact evaluation [Data set]. Retrieved from https://microdata.worldbank.org/index.php/catalog/3499

Font SA, & Cage J. (2018). Dimensions of physical punishment and their associations with children's cognitive performance and school adjustment. Child Abuse & Neglect, 75, 29–40. 10.1016/j.chiabu.2017.06.008 [PubMed: 28743493]

Foster EM (2010). Causal inference and developmental psychology. Developmental Psychology, 46, 1454–1480. 10.1037/a0020204 [PubMed: 20677855]

Gershoff ET (2002). Corporal punishment by parents and associated child behaviors and experiences: A meta-analytic and theoretical review. Psychological Bulletin, 128, 539–579. 10.1037/0033-2909.128.4.539 [PubMed: 12081081]

Gershoff ET (2013). Spanking and child development: We know enough now to stop hitting our children. Child Development Perspectives, 7, 133–137. 10.1111/cdep.12038 [PubMed: 24039629]

Gershoff E. (2016). Should parents' physical punishment of children be considered a source of toxic stress that affects brain development? Family Relations, 65, 151–162. 10.1111/fare.12177

Gershoff ET, Goodman GS, Miller-Perrin CL, Holden GW, Jackson Y, & Kazdin AE (2018). The strength of the causal evidence against physical punishment of children and its implications for parents, psychologists, and policymakers. American Psychologist, 73, 626–638. 10.1037/amp0000327

Gershoff ET, & Grogan-Kaylor A. (2016). Spanking and child outcomes: Old controversies and new meta-analyses. Journal of Family Psychology, 30, 453–469. 10.1037/fam0000191 [PubMed: 27055181]

Gershoff ET, Lansford JE, Sexton HR, Davis-Kean P, & Sameroff AJ (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. 10.1111/j.1467-8624.2011.01732.x [PubMed: 22304526]

Gershoff ET, Lee SJ, & Durrant JE (2017). Promising intervention strategies to reduce parents' use of physical punishment. Child Abuse & Neglect, 71, 9–23. 10.1016/j.chiabu.2017.01.017 [PubMed: 28162793]

Gershoff ET, Sattler KMP, & Ansari A. (2018). Strengthening causal estimates for links between spanking and children's externalizing behavior problems. Psychological Science, 29, 110–120. 10.1177/0956797617729816 [PubMed: 29106806]

Global Initiative to End Corporal Punishment of Children. (2019). Progress. Retrieved from https://endcorporalpunishment.org/countdown/

Gunnar MR, Hostinar CE, Sanchez MM, Tottenham N, & Sullivan RM (2015). Parental buffering of fear and stress neurobiology: Reviewing parallels across rodent, monkey, and human models. Social Neuroscience, 10, 474–478. 10.1080/17470919.2015.1070198 [PubMed: 26234160]

Hainmueller J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. Political Analysis, 20, 25–46. 10.1093/pan/mpr025

Hamadani JD, Tofail F, Hilaly A, Huda SN, Engle P, & Grantham-McGregor SM (2010). Use of family care indicators and their relationship with child development in Bangladesh. Journal of Health, Population, and Nutrition, 28, 23–33. 10.3329/jhpn.v28i1.4520

Horn JL, & Trickett PK (1998). Community violence and child development: A review of research. In Trickett PK & Schellenbach CJ (Eds.), Violence against children in the family and the community (pp. 103–138). Washington, DC: American Psychological Association. 10.1037/10292-004

ICBF. (2017). Lineamiento técnico para la atención de niños, niñas y adolescentes con sus derechos inobservados, amenazados o vulnerados por causa de la violencia [Technical guideline for the care of children and adolescents with threatened or violated rights due to exposure to violence]. Bogotá, Colombia: Instituto Colombiano de Bienestar Familiar.

Jurado Castro V, & Rebolledo-Cobos R. (2017). Análisis de las escalas para la evaluación del desarrollo infantil usadas en América: Una revisión de literatura [Analysis of the evaluation scales for child development in América: A literature review]. Revista Movimiento Científico, 10, 72–82. 10.33881/2011-7191.mct.10206

Lansford JE, Cappa C, Putnick DL, Bornstein MH, Deater-Deckard K, & Bradley RH (2017). Change over time in parents' beliefs about and reported use of corporal punishment in eight countries with and without legal bans. Child Abuse & Neglect, 71, 44–55. 10.1016/j.chiabu.2016.10.016 [PubMed: 28277271]

Larzelere RE, Gunnoe ML, & Ferguson CJ (2018). Improving causal inferences in meta-analyses of longitudinal studies: Spanking as an illustration. Child Development, 89, 2038–2050. 10.1111/cdev.13097 [PubMed: 29797703]

Ma J, Grogan-Kaylor A, & Lee SJ (2018). Associations of neighbour-hood disorganization and maternal spanking with children's aggression: A fixed-effects regression analysis. Child Abuse & Neglect, 76, 106–116. 10.1016/j.chiabu.2017.10.013 [PubMed: 29100038]

Maguire-Jack K, Gromoske AN, & Berger LM (2012). Spanking and child development during the first 5 years of life. Child Development, 83, 1960–1977. 10.1111/j.1467-8624.2012.01820.x [PubMed: 22860622]

McCoy DC (2013). Early violence exposure and self-regulatory development: A bioecological systems perspective. Human Development, 56, 254–273. 10.1159/000353217

Miller P, Henry D, & Votruba-Drzal E. (2016). Strengthening causal inference in developmental research. Child Development Perspectives, 10, 275–280. 10.1111/cdep.12202

Ministerio de Salud y Profamilia. (2016). Encuesta Nacional de Demografía y Salud–Tomo 1 [National demographic and health survey - Volume 1]. Bogotá, Colombia: Author.

Okuzono S, Fujiwara T, Kato T, & Kawachi I. (2017). Spanking and subsequent behavioral problems in toddlers: A propensity scorematched, prospective study in Japan. Child Abuse & Neglect, 69, 62–71. 10.1016/j.chiabu.2017.04.002 [PubMed: 28448815]

Pace GT, Lee SJ, & Grogan-Kaylor A. (2019). Spanking and young children's socioemotional development in low- and middle-income countries. Child Abuse & Neglect, 88, 84–95. 10.1016/j.chiabu.2018.11.003 [PubMed: 30448642]

Patterson GR (1982). Coercive family process. Eugene, OR: Castalia.

Radloff LS. (1977). The CES-D Scale: A self-report depression scale for research in the general population. Applied Psychological Measurement, 1, 385–401. 10.1177/014662167700100306

Rubio-Codina M, Attanasio O, Meghir C, Varela N, & Grantham-McGregor S. (2015). The socioeconomic gradient of child development: Cross Sectional evidence from children 6–42 months in Bogota. The Journal of Human Resources, 50, 464–483. 10.3368/jhr.50.2.464

Sameroff A. (2010). A unified theory of development: A dialectic integration of nature and nurture. Child Development, 81, 6–22. 10.1111/j.1467-8624.2009.01378.x [PubMed: 20331651]

Save the Children. (2017). Stolen childhoods: End of childhood report 2017. Retrieved from https://resourcecentre.savethechildren.net/library/stolen-childhoods-end-childhood-report-2017.

Sege RD, Siegel BS, & Council on Child Abuse and Neglect and the Committee on Psychosocial Aspects of Child and Family Health. (2018). Effective discipline to raise healthy children. Pediatrics, 142, e20183112. 10.1542/peds.2018-3112

Sheu YS, Polcari A, Anderson CM, & Teicher MH (2010). Harsh corporal punishment is associated with increased T2 relaxation time in dopamine-rich regions. NeuroImage, 53, 412–419. 10.1016/j.neuroimage.2010.06.043 [PubMed: 20600981]

Shonkoff JP, Boyce WT, & McEwen BS (2009). Neuroscience, molecular biology, and the childhood roots of health disparities: Building a new framework for health promotion and disease prevention. Journal of the American Medical Association, 301, 2252–2259. 10.1001/jama.2009.754 [PubMed: 19491187]

Tomoda A, Suzuki H, Rabi K, Sheu Y-S, Polcari A, & Teicher MH (2009). Reduced prefrontal cortical gray matter volume in young adults exposed to harsh corporal punishment. NeuroImage, 47(Suppl. 2), T66–T71. 10.1016/j.neuroimage.2009.03.005 [PubMed: 19285558]

Wang MT, & Kenny S. (2014). Parental physical punishment and adolescent adjustment: Bidirectionality and the moderation effects of child ethnicity and parental warmth. Journal of Abnormal Child Psychology, 42, 717–730. 10.1007/s10802-013-9827-8 [PubMed: 24384596]

Wolke D, Skuse D, & Mathisen B. (1990). Behavioral style in failure-to-thrive infants: A preliminary communication. Journal of Pediatric Psychology, 15, 237–254. 10.1093/jpepsy/15.2.237 [PubMed: 2374078]

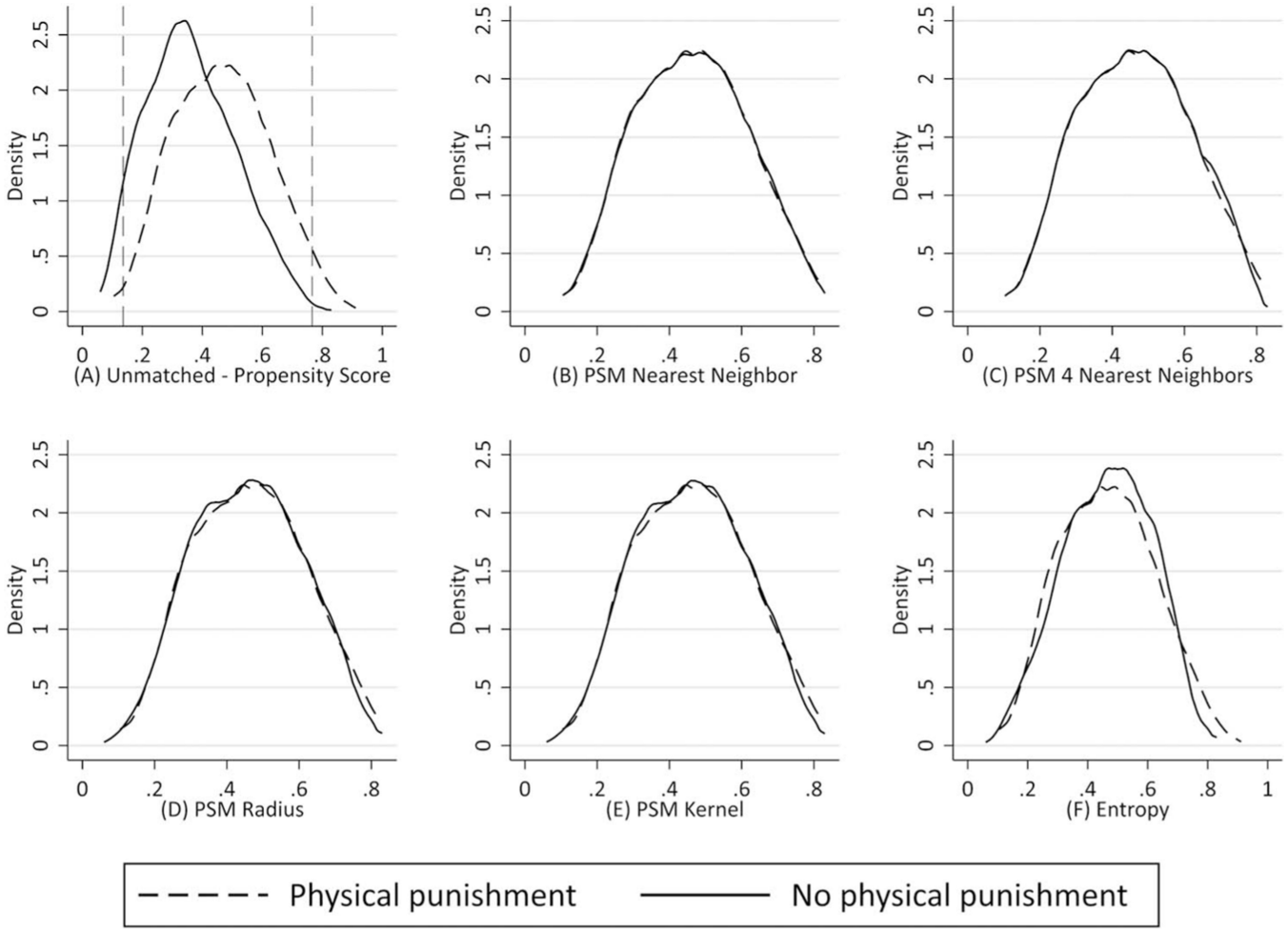

Figure 1.

Density of propensity scores according to exposure to physical punishment before and after matching techniques.

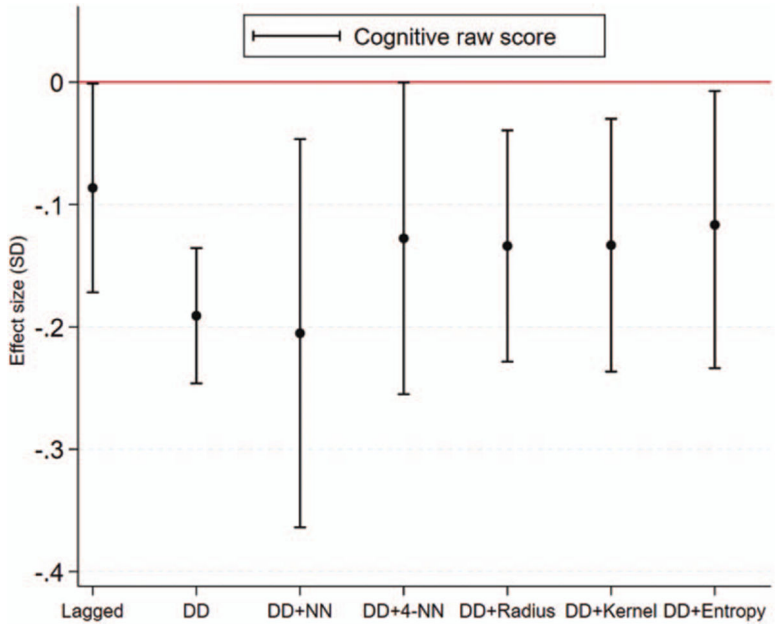

Figure 2.

Summary of results across methods characterizing links between physical punishment (T1) and children's cognitive development (T2). DD = difference in differences; NN = nearest neighbor matching; 4-NN = four nearest neighbors matching; clustered standard errors were used for lagged, DD, and entropy matching and bootstrapped standard errors with 100 repetitions for propensity score matching (PSM); bars represent 95% confidence intervals. See the online article for the color version of this figure.

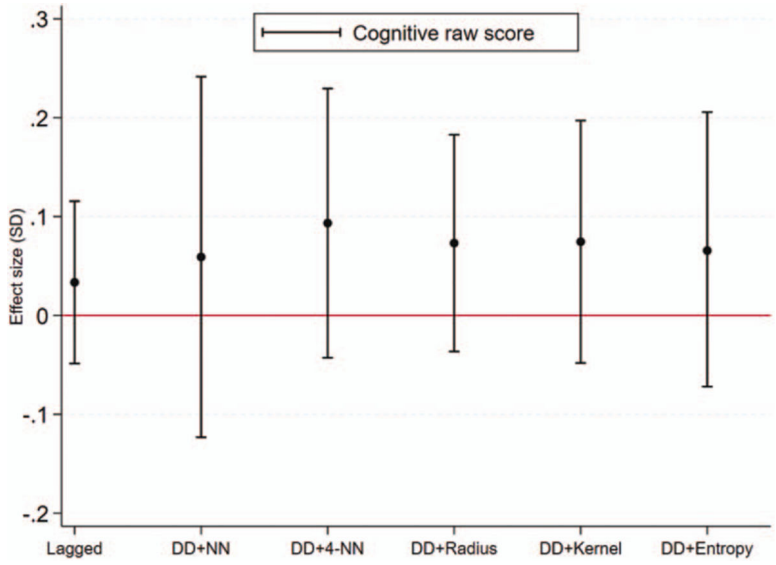

Figure 3.

Results of falsification test predicting T1 cognitive skills from T2 physical punishment in reversed time. (See Table S6 for details). DD = difference in differences; NN = nearest neighbor match; 4-NN = four nearest neighbors match; clustered standard error for Lagged, DD, and Entropy matching and bootstrapped standard errors with 100 repetitions for propensity score matching (PSM); bars represent 95% confidence intervals. See the online article for the color version of this figure.

Table 1

Sample Characteristics in T1 (N = 1,167)

Variable	M	SD	Observed minimum	Observed maximum
Independent variable				
Physical punishment (ever) T1	0.41	0.49	0	1
Dependent variable				
Cognitive skills T2 (Bayley-III raw score)	72.09	4.52	37	84
Covariates (all T1)				
Child characteristics				
Age in months	17.79	3.74	9	26
Sex (1 = boy; 0 = girl)	0.51	0.50	0	1
Hemoglobin concentration (g/dL)	11.36	1.30	6.70	15.10
Weight (kg)	10.37	1.43	6.2	17
Height (cm)	79.06	4.47	66.90	91.80
Child development (Bayley-III raw scores)				
Cognitive skills	51.83	7.39	28	72
Receptive language	20.25	4.98	6	36
Expressive language	20.21	6.33	3	44
Fine motor skills	34.39	3.95	17	50
Gross motor skills	50.53	6.82	26	67
Temperament				
Activity level	3.22	0.76	1	5
Emotional tone	3.33	0.94	1	5
Cooperativeness	3.38	0.97	1	5
Mother characteristics				
Age in years	26.27	6.61	14	47
Highest level of education				
Primary	0.35	0.48	0	1
Secondary or higher	0.65	0.48	0	1
Currently attends school	0.10	0.30	0	1
Depressive symptoms	9.04	5.47	0	28

Variable	M	SD	Observed minimum	Observed maximum
Index of disability	0.18	0.39	0	1
Maternal beliefs regarding child development (1 = yes; 0 = no)				
Children's intelligence is fixed	0.35	0.48	0	1
Knowing words improves learning	0.88	0.33	0	1
Playing increases achievement	0.87	0.34	0	1
Children start to talk according to nature	0.74	0.44	0	1
School achievement determines earnings	0.60	0.49	0	1
Playing alone is good	0.32	0.47	0	1
Better to wait until telling stories	0.48	0.50	0	1
Praising makes the child overly confident	0.46	0.50	0	1
A busy parent should spend time playing	0.80	0.40	0	1
Home environment				
Father lives at home	0.67	0.47	0	1
A household member died recently	0.02	0.16	0	1
Family wealth	0.04	1.63	−10.87	3.19
Stimulating materials	3.59	1.59	0	9
Cognitive stimulation	3.05	1.61	0	7
Municipality characteristics				
Population (in 1,000s)	25.70	13.28	6.53	72.96
Number of resources	45.33	31.84	10	190
Number of shocks	1.76	1.55	0	6
Multidimensional poverty index (2005)	63.19	11.25	32.39	87.76
Unsatisfied basic needs index (2011)	34.07	13.70	11.14	100
Average Saber 11 test score	242.31	10.75	212.621	279.511
FARC guerrillas (1 = present; 0 = not present)	0.80	0.40	0	1
Homicide rate per 100,000	32.88	20.77	0	95.03
Theft rate per 100,000	271.80	289.28	0	1492.02
Terrorism rate per 100,000	9.74	16.75	0	71.57

Table 2

Results From Lagged Dependent Variable Models Predicting T2 Standardized Cognitive Skills

Outcome: Cognitive skills (*SD*) at T2	**(1)**	**(2)**	**(3)**	**(4)**
Physical punishment at T1	−0.22***	−0.08†	−0.09*	−0.09*
	(0.06)	(0.04)	(0.04)	(0.04)
Cognitive skills (SD) at T1		0.91**	0.53**	0.51**
		(0.09)	(0.09)	(0.09)
Covariates	No	No	Yes	Yes
Number of children	1,167	1,167	1,167	1,167
R-squared	0.01	0.40	0.44	0.45

Note. Clustered-standard errors in parentheses Table S3 in the online supplemental materials presents coefficients for all covariates. Covariates included in Model 2: child age in months and sex. Covariates included in Model 3: child age in months and sex, receptive and expressive language, fine and gross motor. Model 4: same as Model 3 plus child activity, emotional tone and cooperativeness during BSID-III implementation, hemoglobin concentration, weight, maternal education, and wealth index.

† $p < .1$.

* $p < .05$.

** $p < .01$.

*** $p < .01$.

Table 3

Results From Difference-in-Differences Models Predicting Cognitive Skills

Outcome: Cognitive skills (*SD*)	**(1)**	**(2)**	**(3)**
Time (T2; α_2)	1.78 **	1.78 **	1.79 **
	(0.02)	(0.02)	(0.02)
Physical punishment (α_1)	0.28 **	0.11 **	0.11 **
	(0.04)	(0.02)	(0.02)
Physical punishment × Time (β)	**−0.20** **	**−0.20** **	**−0.19** **
	(0.03)	**(0.03)**	**(0.03)**
Covariates	No	Yes	Yes
Observations	2,334	2,334	2,334
Number of children	1,167	1,167	1,167
Overall R-squared	0.74	0.88	0.90

Note. Clustered-standard errors in parentheses Table S4 in the online supplemental materials presents coefficients for all covariates. All Covariates are at T1. Covariates included in (2): child age in months, sex, receptive and expressive language, fine and gross, motors. In (3): the same as (2) and child activity, emotional tone and cooperativeness during BSID-III implementation, hemoglobin, weight, height, maternal depression, age, education, disability index, stimulating materials at home, index of stimulation, maternal beliefs, and municipality characteristics.

** $p < .01$.

Table 4

Results From Difference-in-Differences-With-Matching Models Predicting Cognitive Skills at T2 From Physical Punishment at T1

Approach	Observed coefficient	Standard error[a]	95% CI	*N* in common support
PSM—NN	**−0.21** *	0.08	[−0.36, −0.05]	1,154
PSM—4 NN	**−0.13** *	0.06	[−0.26, −0.00]	1,154
PSM—Radius	**−0.13** **	0.05	[−0.23, −0.04]	1,154
PSM—Kernel	**−0.13** *	0.05	[−0.24, −0.03]	1,154
Entropy	**−0.12** *	0.06	[−0.23, −0.00]	1,167

Note. PSM = propensity score matching. In bold: Double difference, which represents the effect of corporal punishment on children's cognitive development.

[a]Bootstrapped standard error with 100 repetitions for PSM and clustered standard error for entropy matching.

* $p < .05$.

** $p < .01$.