

Comparing Child Outcomes of Physical Punishment and Alternative Disciplinary Tactics: A Meta-Analysis

Robert E. Larzelere^{1,2} and Brett R. Kuhn¹

This meta-analysis investigates differences between the effect sizes of physical punishment and alternative disciplinary tactics for child outcomes in 26 qualifying studies. Analyzing differences in effect sizes reduces systematic biases and emphasizes direct comparisons between the disciplinary tactics that parents have to select among. The results indicated that effect sizes significantly favored *conditional* spanking over 10 of 13 alternative disciplinary tactics for reducing child noncompliance or antisocial behavior. *Customary* physical punishment yielded effect sizes equal to alternative tactics, except for one large study favoring physical punishment. Only *overly severe* or *predominant* use of physical punishment compared unfavorably with alternative disciplinary tactics. The discussion highlights the need for better discriminations between effective and counterproductive use of disciplinary punishment in general.

KEY WORDS: children; parenting; discipline; punishment; spanking.

Uncertainty about the effects of physical punishment on children has persisted despite decades of research. Two major perspectives have emerged recently. The first is an unconditional anti-spanking perspective, advanced by both social scientists (Gershoff, 2002; Straus, 2001) and advocacy groups (EPOCH-Worldwide, 2004). In response, at least 13 countries have passed laws banning all physical punishment by parents (EPOCH-Worldwide, 2004).

The second perspective, which has been called the conditional-spanking perspective (Benjet & Kazdin, 2003), has attempted to identify conditions under which spanking may be beneficial or at least not detrimental to children. The conditional-spanking perspective emphasizes the parenting context and manner of implementation, which may distinguish effective from counterproductive uses of punishment more than its form (e.g., physical or nonphysical). In one sense the disciplinary ac-

tions of parents in most cultures have reflected a conditional-spanking perspective until recently. In 1994–1995, for example, 94% of American parents and 52% of Canadian parents of 3- and 4-year-olds reported using physical punishment at least occasionally (Larzelere, 2004; Straus & Stewart, 1999). The conditional-spanking perspective holds that spanking should be investigated under the conditions for which parents have considered it advisable before imposing a spanking ban on parents (Bauman & Friedman, 1998; Baumrind, Larzelere, & Cowan, 2002; Eysenck, 1993; Friedman & Schonberg, 1996b; Larzelere, Baumrind, & Polite, 1998).

Two recent literature reviews from these two perspectives did little to resolve the issue. Gershoff's (2002) meta-analysis concluded that physical punishment was linked positively to immediate compliance, but negatively with 10 other outcomes in children and families. In a qualitative review, Larzelere (2000) concluded that causal evidence showed that nonabusive spanking of 2–6-year-olds produced more beneficial than detrimental child outcomes when it was used to enforce milder disciplinary tactics such as reasoning or time-out, especially in subcultural groups that support its use.

¹Psychology Department, Munroe-Meyer Institute, University of Nebraska Medical Center, Nebraska.

²Address all correspondence to Robert E. Larzelere, Psychology Department, MMI, 985450 Nebraska Medical Center, Omaha, Nebraska 68198-5450; e-mail: rlarzelere@unmc.edu.

Benjet and Kazdin (2003) recently compared the two reviews and concluded, "A top priority for research on spanking would seem to be a comparison of spanking with alternative procedures that already have considerable evidence in their behalf" (p. 215). The current meta-analysis attempts to address this priority by investigating the studies included in either review that examined one or more alternative disciplinary tactics in addition to physical punishment. It also investigates several methodological problems that could explain the discrepant conclusions from the two reviews (Benjet & Kazdin, 2003).

To provide a context for this meta-analysis, we briefly summarize the methodological problems that have hindered definitive conclusions about physical punishment. We then clarify why a meta-analysis using differences in effect sizes between physical punishment and disciplinary alternatives can reduce these methodological problems.

Methodological Issues

The spanking controversy persists largely because pervasive methodological problems have permitted a wide range of interpretations. These problems include predominantly correlational research; failing to discriminate among nonabusive, customary, and overly severe use of physical punishment; measuring disciplinary practices and child outcomes from the same information source; and failing to rule out plausible alternative explanations.

The strongest evidence against physical punishment in Gershoff's (2002) thorough meta-analysis consisted of longitudinal correlations, i.e., zero-order correlations between physical punishment and subsequent child outcomes. Although such correlations are consistent with a causal effect (Smith, 2002), their pattern is typical of most corrective interventions (Larzelere, Kuhn, & Johnson, 2004). In post-treatment comparisons, recipients of corrective interventions will compare poorly to those not needing such interventions, whether the intervention is delivered by physicians (e.g., radiation treatment), educators (Head Start), psychologists (marital counseling), or parents (punishment).

Consider radiation treatment as an example. Patients who received radiation treatment last year are more likely to experience cancer-related symptoms this year than those who did not receive (or need) radiation treatment. Longitudinal zero-order correlations would indicate that radiation treatment is

associated with increased cancer-related symptoms. Of course, the initial presenting problem (cancer) is the causal factor underlying that correlation because it leads to both the corrective intervention (radiation treatment) and the subsequent outcome (cancer). Consequently, zero-order longitudinal correlations cannot discriminate effective corrective interventions from those that are counterproductive.

Second, most of the research on physical punishment "lumps" together nonabusive and customary punishment with overly severe forms of physical punishment. For example, 65% of the studies in Gershoff's (2002) meta-analysis included overly severe physical punishment in their measure, according to Baumrind et al. (2002). Examples ranged from vaguely defined "punitive discipline" (6% of the studies), composite measures of the frequency and severity of physical punishment (29%), and the inclusion of extreme violence (31%), such as slapping in the face (seven studies), beating up (three studies), or hitting with a fist and causing bruises and cuts (one study).

Third, many studies of disciplinary tactics have based the antecedent and consequent variables on the same source of information. Typically, mothers reported both their disciplinary tactics and their child's behavior. In retrospective studies, grown children reported both their current functioning and the disciplinary tactics they received earlier in life. This same-source bias has been shown to inflate associations between disciplinary tactics and adverse outcomes (Yarrow, Campbell, & Burton, 1968).

Finally, plausible alternative explanations of the data on physical punishment have not been ruled out, resulting in widely discrepant explanations for the varied outcomes across studies. Consider the strongest evidence of the effectiveness of spanking. Four small-randomized clinical studies found that spanking was effective in reducing defiance in clinically oppositional 2–6-year-olds (Bean & Roberts, 1981; Roberts; Day & Roberts, 1983; Roberts, 1988; Roberts & Powers, 1990). The difference in effect sizes between those four randomized studies (mean $d = 1.21$) and the 113 non-randomized studies (mean $d = -.35$) in Gershoff (2002) approached the largest difference ever found in a meta-analysis (Lipsey & Wilson, 1993). This difference could be explained by one or more of the following confounded interpretations. Compared to the non-randomized studies, Roberts' four randomized studies (1) had causally stronger evidence, (2) limited spanking to two open-handed swats under the supervision of a clinical

psychologist, (3) used spanking only to enforce compliance with time-out, (4) applied only to children from 2 to 6 years of age who (5) were clinically referred for oppositional behavior problems, and (6) focused on decreases in defiance in the clinic as the primary outcome. Whereas advocates of the anti-spanking viewpoint consider the type of outcome (short-term compliance) to be the crucial distinction (Gershoff, 2002; Straus, 2001), conditional-spanking researchers emphasize the stronger causal evidence, the specific conditions in the randomized studies (e.g., the child's age, the discipline situation), and the way in which spanking was implemented (Baumrind et al., 2002; Larzelere, 2000).

Although these four methodological problems are often acknowledged, the extent to which they undermine research conclusions has received insufficient attention. Suppose radiation treatment were studied in the same way that researchers have investigated physical punishment. Borrowing statistics from Gershoff's (2002) thorough meta-analysis, two-thirds (65%) of studies of radiation treatment would have included excessive dosages of radiation (Baumrind et al., 2002), 58% would have been cross-sectional studies, and only 4% would have taken into consideration the presence or severity of cancer. Would it be surprising that patients who received radiation treatment last year had higher rates of cancer both last year and this year, compared to those who did not receive (or need) radiation? A meta-analysis of radiation treatment using predominantly correlational studies would come to the same conclusions as Gershoff's (2002) meta-analysis, specifically that radiation treatment is consistently linked to detrimental outcomes. As aptly noted by Straus (2001), valid causal conclusions require controlling for the effects of initial child misbehavior. Otherwise, initial child misbehavior may lead to more disciplinary tactics as well as worse child outcomes, which would account for the associations found by Gershoff (2002).

Rationale for a Meta-Analysis of Differential Effect Sizes

This meta-analysis attempts to reduce these pervasive methodological problems by (1) distinguishing among four types of physical punishment, (2) basing effect sizes on each study's strongest methodological evidence whenever possible, and (3) analyzing *differential* effect sizes between physical punishment and alternative disciplinary tactics.

To address the "lumping" problem, we distinguish among conditional spanking, customary physical punishment, overly severe physical punishment, and predominant use of physical punishment. *Conditional* spanking (as labeled by Benjet & Kazdin, 2003) refers to spanking under the limited conditions that have been associated with better child outcomes (e.g., spanking when a 2–6-year-old refuses to comply with time out). The purpose of distinguishing this category is to determine whether spanking is associated with better outcomes than alternative tactics even under ideal conditions. *Customary* physical punishment represents the manner in which parents typically use physical punishment. The purpose of this category is to investigate whether typical use of physical punishment is associated with better or worse outcomes than alternative tactics. *Overly severe* physical punishment includes the use of excessive force, hitting with an object, or slapping in the face (Baumrind et al., 2002). Finally, *predominant* usage indicates that physical punishment is the parent's primary disciplinary method, i.e., it is preferred over milder disciplinary tactics.

This meta-analysis bases effect sizes on the findings from each study that are methodologically strongest. For example, our effect sizes are based on results that take initial child misbehavior into account from distinct sources of information, whenever possible. This choice contrasts with Gershoff's (2002) decision to base effect sizes on correlations for the sake of consistency, ignoring methodologically stronger findings in several studies.

Finally, this meta-analysis estimates differences in the effect sizes of physical punishment vs. alternative disciplinary tactics, using identical methods within the same study. If the apparently detrimental child outcomes reflect causal effects unique to physical punishment, then the effect sizes of physical punishment should compare poorly to the effect sizes of alternative disciplinary tactics. On the other hand, if detrimental child correlates of physical punishment represent methodological artifacts, then the effect sizes of alternative disciplinary tactics should appear equally detrimental.

A methodology for analyzing differences between effect sizes is already well established for randomized studies. It is based on the differential effect size contrasting post-treatment outcomes from a treatment and a control group. For the usual effect size measure (*d*), this is the same as calculating an effect size for each group (e.g., improvement from pre-to-post) and then using the difference between

those two effect sizes. This equality is based on two assumptions. First, the treatment and control group must have identical pre-treatment scores, which randomization guarantees in the long run.³ The second assumption is that the effect sizes for the treatment and control groups are based on the same standard deviation. When these assumptions apply, typical meta-analyses of randomized clinical trials can be considered equivalent to analyses of differences between effect sizes. The only distinction in our meta-analysis is that it compares two treatments (disciplinary tactics) with each other rather than treatment and control groups.

The major advantage of analyzing differences between effect sizes, however, is for non-randomized studies, which dominate this literature. Causal conclusions can be supported from correlational studies only to the extent that plausible alternative interpretations have been ruled out (Larzelere et al., 2004; Shadish, Cook, & Campbell, 2002). This principle applies to individual studies as well as to meta-analyses. For example, an alternative explanation for the positive correlation between physical punishment and subsequent antisocial behavior is that the child's initial antisocial behavior may increase both the frequency of physical punishment and subsequent antisocial behavior. Just as individual studies control for this possibility by using initial child misbehavior as a covariate, this meta-analysis uses differences between effect sizes to control for initial child misbehavior.

A second advantage of analyzing differences between effect sizes is that they allow researchers to directly compare realistic disciplinary choices. Basing effect sizes on simple associations between a disciplinary tactic and a child outcome implicitly compares parents who use that disciplinary tactic with those who do not use it. Instead of choosing between a given disciplinary tactic and doing nothing, parents typically choose between two or more alternative disciplinary responses (Ritchie, 1999). Differences in effect sizes are better suited for such comparisons.

In summary, this meta-analysis uses differences between effect sizes to control for confounds that influence all disciplinary tactics, e.g., selection bias due to initial child misbehavior. It falls short of being causally definitive, however, because it rules out only some plausible interpretations of the underlying empirical evidence. This strategy is a substantial

improvement over typical meta-analytic methods for correlational data because it controls for important confounds and rules out alternative interpretations associated with them. At the very least, the current meta-analysis can determine whether the correlationally based effect sizes are uniquely detrimental for physical punishment, are more detrimental for some disciplinary tactics than others, or are equally detrimental for all disciplinary tactics. Making these distinctions is a crucial step toward designing more causally informative studies in the future. The results also have important implications for how physical punishment should be used, if at all, and which alternative disciplinary tactics might be used instead.

METHOD

Literature Selection

Research studies were selected for this meta-analysis from recent reviews by Gershoff (2002) and Larzelere (2000). Both reviews attempted to be exhaustive within their inclusion criteria for at least the previous 26 years. Additional selection criteria include the following: (1) The study must have investigated one or more recommended alternative disciplinary tactics as well as physical punishment, using similar research methods. (2) The children had to average less than 13 years old at the time of the discipline. Most retrospective studies were excluded because they pertained to physical punishment of teenagers, based on the finding that retrospective reports of physical punishment correlated most highly with mothers' reported physical punishment at 12–14 years old (Stattin, Janson, Klackenberg-Larsson, & Magnusson, 1995). To be included, retrospective surveys had to ask specifically for disciplinary tactics at a younger age. (3) Selected studies had to investigate at least one child outcome, excluding studies that investigated only parental outcomes.

These criteria yielded 26 studies that investigated physical punishment and one or more alternative tactics, summarized in Table I. Only eight of these studies were included in both previous reviews. Eleven studies from Gershoff's (2002) meta-analysis were excluded from Larzelere's (2000) review because they were cross-sectional (seven studies) or used overly broad measures of punishment (three studies). One other study was incorrectly excluded from Larzelere's (2000) review, because it did specify a younger age in its retrospective survey

³When pre-test scores differ, relative pre-post gains provide a fairer comparison than post-treatment differences.

Table 1. Continued

Study	Age, Gender (Parent) ^a	Sample	N	Discipline tactic ^b	Outcome	Effect size ^c (d)	Basis of effect size & discrepancies from Gershoff (2002)
			9	Restraint back-up for TO	TO success	-.22	Same
Within-subject sequential analyses							
Larzelere et al. (1996)	2-3 B, G (M)	Volunteers	38	Slap hand or spank, whether reasoning was also used or not (customary PP, contrasted with next two tactics)	Delay until next recurrence of disobedience, compared to typical delays for that child	-.05	Deviations from participants' mean delays, compared with "other" (i.e., no punishment or reasoning). Could not replicate Gershoff
				Reasoning, whether used with physical or nonphysical punishment or not	Delay until disobedience recurrence	.01	Same
				Nonphysical punishment (time-out or privilege removal), whether used with reasoning or PP or not	Delay until disobedience recurrence	-.02	Same
				Reasoning & PP (no nonphysical punishment; conditional PP combined this with Reasoning & PP & nonphysical punishment)	Delay until disobedience recurrence	-.02	Same
				Reasoning & nonphysical punishment & PP (part of conditional PP)	Delay until disobedience recurrence	.17	Same
				Reasoning alone (no physical or nonphysical punishment)	Delay until disobedience recurrence	.06	Same
				Nonphysical punishment alone (time-out or privilege removal; no reasoning or PP)	Delay until disobedience recurrence	.02	Same
				Reasoning & nonphysical punishment (no PP)	Delay until disobedience recurrence	-.08	Same

Physical Punishment vs. Alternative Tactics

Author (Year)	Sample	Source	Intervention	Outcome	Effect Size	Notes
Ritchie (1999)	3 B, G (M)	Volunteers from birth records	Spank (conditional PP)	Immediate reduction in probability of defiance	.97	From immediately prior probability of defiance compared to immediately subsequent probability of defiance. Study not in Gershoff
			Reason or offer alternatives	Drop in defiance	-.02	Same
			Threaten or verbal power assertion	Drop in defiance	.08	Same
			Privilege removal	Drop in defiance	-.33	Same
			Time-out	Drop in defiance	.60	Same
			No response (ignore)	Drop in defiance	.02	Same
			Physical power assertion	Drop in defiance	.45	Same
			Spank (customary PP)	Immediate reduction in "physical" or passive noncompliance	.07	From immediately prior probability of two noncompliance types compared to their immediately subsequent probability
			Reason or offer alternatives	Reduction in noncompliance	.18	Same
			Threaten or verbal power assertion	Reduction in noncompliance	.02	Same
			Privilege removal	Reduction in noncompliance	.24	Same
			Time-out	Reduction in noncompliance	.16	Same
			No response (ignore)	Reduction in noncompliance	.33	Same
Physical power assertion	Reduction in noncompliance	.20	Same			
Correlational sequential analyses Chapman and Zahn-Waxler (1982)	10-29 mos., B, G (M)	Volunteers	Physical coercion (PP or restraint) without reasoning (customary PP)	Immediate compliance	.09	Compared to the overall compliance rate. Could not replicate Gershoff
			Physical coercion & reasoning (conditional PP)	Immediate compliance	.02	Compared to the overall compliance rate
			Reasoning (with or without verbal prohibition)	Immediate compliance	-.22	Same
			Verbal prohibition	Immediate compliance	-.15	Same
			Love withdrawal (including ignoring and time-out) plus any of above tactics	Immediate compliance	.40	Same
Minton, Kagan, and Levine (1971)	27 mos. B, G (M)	Volunteers	PP as proportion of observed misbehavior (predominant PP)	Disobedience requiring maternal reprimand	-.55	Averaged correlations for boys and girls. Same as Gershoff

Table 1. Continued

Study	Age, Gender (Parent) ^a	Sample	N	Discipline tactic ^b	Outcome	Effect size ^c (d)	Basis of effect size & discrepancies from Gershoff (2002)
<i>Antisocial behavior</i>							
Statistically controlled longitudinal studies							
Larzelere and Smith (2000)	6-9 at T1, 8-11 at T2; B, G (M)	National sample of young mothers	785	Explanations of reprimands as proportion of misbehavior Frequency spanked in past week (customary PP)	Observed disobedience Antisocial behavior 2 years later	.44 -.23	Averaged correlations for boys and girls Mean antisocial for 1 or more times per week vs. no use of the disciplinary tactic, controlling for externalizing problems at Time 1, five other variables and 6 interactions of these variables with the disciplinary tactic. Unpublished. not in Gershoff
			785	Frequency privileges removed in past week	Antisocial behavior later	-.21	Same
			785	Frequency grounded in past week	Antisocial behavior later	-.20	Same
			771	Frequency allowance removed in past week	Antisocial behavior later	-.10	Same
			785	Frequency sent to room in past week	Antisocial behavior later	-.18	Same
Larzelere et al. (1998)	2-3 at T1, 4 at T2; B, G (M)	Volunteers	38	Slap hand or spank (PP) without reasoning, as proportion of misbehavior incidents (averaged with Reasoning & PP for predominant PP)	Disruptive behavior 20 months later	.41	Mean partial correlation of proportional usage with subsequent disruptive behavior, controlling for initial disruptive behavior. Not in Gershoff (multiple reports from same study)
				Reasoning & PP (proportional use; part of predominant PP)	Disruptive behavior 20 months later	-.32	Same
				Reasoning without PP or nonphysical punishment	Disruptive behavior later	-.80	Same
				Nonphysical punishment (time-out or privilege removal)	Disruptive behavior later	.20	Same
				Reasoning & nonphysical punishment	Disruptive behavior later	.10	Same

Physical Punishment vs. Alternative Tactics

9

Study	Sample	Intervention	Comparison	Outcome	Effect Size	Notes																
Within-subject sequential analyses Larzelere et al. (1996)	Volunteers	Slap hand or spank (PP), whether reasoning was also used or not (customary PP, contrasted with next two tactics)	Reasoning, whether used with physical or nonphysical punishment or not	Delay until next recurrence of fighting, compared to typical delays for that child	.08	Deviations from participants' mean delays, compared with "other" (i.e., no punishment or reasoning). Gershoff did not include this outcome																
					Reasoning, whether used with physical or nonphysical punishment or not	Delay until fighting recurrence	.01	Same														
							Nonphysical punishment (time-out or privilege removal), whether used with reasoning or PP or not	Delay until fighting recurrence	.36	Same												
									Reasoning & PP (no nonphysical punishment; conditional PP combined this with Reasoning & PP & nonphysical punishment)	Delay until fighting recurrence (deviation)	.07	Same										
											Reasoning & nonphysical punishment & PP (part of conditional PP)	Delay until fighting recurrence	.81	Same								
													Reasoning alone (no physical or nonphysical punishment)	Delay until fighting recurrence	-.09	Same						
															Nonphysical punishment alone (time-out or privilege removal; no reasoning or PP)	Delay until fighting recurrence	.26	Same				
																	Reasoning & nonphysical punishment (no PP)	Delay until fighting recurrence	.63	Same		
																			Uncontrolled longitudinal studies McClelland and Pilon (1983)	Need for Power 26 years later	.42	Mean of correlations with Need for Power for males and females. Study not in Gershoff
																					Reasoning severity (severe PP)	Need for Power
Privilege removal	Need for Power	.00	Non-significant <i>r</i>																			

Uncontrolled longitudinal studies
McClelland and Pilon (1983)

Kindergarten sample

Extent of PP, combining frequency and severity (severe PP)

Need for Power 26 years later

Mean of correlations with Need for Power for males and females. Study not in Gershoff

Non-significant *r*
Non-significant *r*

Table I. Continued

Study	Age, Gender (Parent) ^a	Sample	N	Discipline tactic ^b	Outcome	Effect size ^c (<i>d</i>)	Basis of effect size & discrepancies from Gershoff (2002)
Sears (1961)	5 at T1, 12 at T2; B, G (M)	Kindergarten sample	160	Love withdrawal Extent of PP, combining severity and frequency (severe PP)	Need for Power Antisocial aggression 7 years later	.00 .14	Non-significant <i>r</i> Correlations with antisocial aggression 7 years later. Gershoff averaged 6 correlations (cross-sectional and with prosocial, ambiguous, and antisocial aggression) Correlations with later antisocial aggression
Yarrow et al. (1968)	4 B, G (M)	Nursery school sample	58	Privilege removal Love withdrawal Use of physical punishment for vignettes about extreme disobedience (conditional PP)	Antisocial aggression Teacher-rated aggression in nursery school 2 months later	-.12 .11 .38	Same Correlation. Gershoff used the correlation of severity of all punishment for aggression with concurrent aggression toward parents (mother-report)
				Use of reasoning from vignettes	Later school aggression	-.28	Correlation
				Use of scolding from vignettes	Later school aggression	-.24	Same
				Use of privilege removal from vignettes	Later school aggression	.38	Same
				Use of isolation from vignettes	Later school aggression	-.18	Same
				Use of diverting from vignettes	Later school aggression	-.47	Same
				Use of love withdrawal from vignettes	Later school aggression	-.24	Same
Retrospective studies Watson (1989)	0-5 at T1; 17 at T2; B, G (M, F)	National Merit Scholarship finalists & average test-takers	2500	Parent-reported spanking and possibly time out before age 6, (customary PP)	Youth-reported hostility & milder ("obloquial") problems	-.09	Correlations with hostility and obloquial problems, using .00 for non-significant <i>r</i> s. Gershoff used only the significant <i>r</i> with one hostility measure
				Privilege removal and assigning extra duties before age 6	Hostility and milder behavior problems	-.13	Correlations with hostility & obloquial problems, using .00 for non-significant <i>r</i> s
Uncontrolled cross-sectional studies Straus and Mouradian (1998)	2-14 B, G (M)	Random sample of two countries	744	How often "spanked, slapped or hit" the child during the past 6 months, controlling for severe out-of-control PP (conditional PP)	Antisocial and impulsive behavior	-.14	<i>F</i> -values for PP, controlling for severe PP, 3 other disciplinary tactics, 4 other variables, and their interactions with PP. Gershoff probably used graphed mean antisocial scores, which could not be compared with alternative tactics

Study	Sample	Sample Size	Design	Findings	Conclusions
<i>Substance abuse</i> Retrospective studies Tennant et al. (1975)	0-14 at T1, M = 23 at T2; B (M, F)	5044	US Army soldiers	Percentage of spankings in which mothers said they "lost it" due to anger (severe PP) How often they used disciplinary reasoning, privilege removal, and time-out during past 6 months	Antisocial and impulsive behavior Antisocial and impulsive behavior
				Spanking (customary PP) Non-contact punishment	-.28 F-values for severe PP, with above controls -.39 F-values for these alternative disciplinary tactics, with above controls
Watson (1989)	0-5 at Time 1; 17 at Time 2. B, G (B)	2500	National Merit Scholarship finalists & average test-takers	Parental report of spanking and possibly time out before age 6 (customary PP) Withdrawal of privileges and assigning extra duties before age 6	Frequent use of hashish, alcohol, amphetamines, and opiates Frequent use of hashish, alcohol, amphetamines, and opiates Youth-reported alcohol usage Youth-reported alcohol usage
					.24 Percentage of most and least frequent users reporting being spanked. Averaged across 4 substances. Not among Gershoff's 11 outcomes -.08 Percentage of most vs. least frequent users reporting receiving this punishment. Averaged across 4 substances -.02 Correlation. Outcome not in Gershoff -.10 Correlation
<i>Conscience & resistance to temptation</i> Uncontrolled longitudinal studies Grinder (1962)	5-6 at T1, 11-12 at T2; B, G (M)	140	Kindergarten sample	Extent of PP, combining severity and frequency (severe PP) Privilege removal Isolation Love withdrawal	Resists temptation in forbidden-toy lab test 6 years later Resists temptation Resists temptation Resists temptation
					-.24 The percentage of children who resisted temptation for high vs. low use of the disciplinary tactic. Could not replicate Gershoff's beneficial effect size .10 Same -.01 Same .37 Same
Uncontrolled cross-sectional studies Aronfreed (1961)	12 B, G (M)	120	Sixth-grade sample	Primarily sensitization (PP & uncontrolled verbal assaults) vs. primarily induction (love withdrawal, encouraging responsibility, and explanations; based on responses to vignettes; predominant PP)	Internal and external motivations for moral corrections in projective story completions 2 x 2 contingency tables with 6 internal and 6 external moral motivations, counting 9 non-significant associations as .00. Gershoff used only the significant association with reparations

Table I. Continued

Study	Age, Gender (Parent) ^a	Sample	N	Discipline tactic ^b	Outcome	Effect size ^c (<i>d</i>)	Basis of effect size & discrepancies from Gershoff (2002)
Burton et al. (1961)	4 B, G (M)	Private nursery school sample	77	Interviewer rating of PP as usual discipline technique (predominant PP)	Resists temptation (lab test)	.35	From 2 × 2 contingency table. Gershoff did not include this outcome
				Reasoning rated as usual technique	Resists temptation	-.08	2 × 2 contingency table
				Scolding rated as usual technique	Resists temptation	.25	Same
				Privilege removal rated as usual technique	Resists temptation	-.63	Same
				Isolation rated as usual technique	Resists temptation	-.12	Same
				Interviewer rating of frequency of spanking, slapping, and shaking (severe PP)	Resists temptation	.00	Table not given, estimated at .00 due to non-significance
				Rated frequency of reasoning	Resists temptation	-.61	2 × 2 contingency table
				Rated frequency of scolding	Resists temptation	.00	Non-significant association
				Rated frequency of privilege removal	Resists temptation	-.34	2 × 2 contingency table.
				Rated frequency of isolation	Resists temptation	.00	Non-significant association
				Rated frequency of love withdrawal	Resists temptation	.33	2 × 2 contingency table
				Interviewer rating of frequency of spanking, slapping, and shaking (severe PP)	Conscience (initial child actions after wrongdoing)	-.62	Gershoff used only this correlation from the study
				Rated frequency of reasoning	Conscience	.36	Correlation
				Rated frequency of scolding	Conscience	.00	Non-significant <i>r</i>
				Rated frequency of privilege removal	Conscience	-.32	Correlation
				Rated frequency of isolation	Conscience	.30	Correlation
				Rated frequency of love withdrawal	Conscience	.00	Non-significant <i>r</i>

Lytton (1977)	2 B (M, F)	Volunteers	90	Maternal and paternal frequencies of PP (customary PP)	Compliance (two measures) and conscience (1)	-.04	Average of betas of 2 parents for 3 outcomes, assuming .00 for non-significant predictors. Gershoff only used the one significant correlation
				Rating of mothers' induction (e.g., explanation of orders)	Compliance and conscience	.06	Non-significant <i>r</i> s, but mostly in the indicated direction, according to the text
				Rating of mothers' verbal psychological punishment (criticism, withdrawal of love)	Compliance and conscience	-.11	Average of betas for three outcomes, assuming .00 for non-significant predictors
				Maternal frequency of love withdrawal	Compliance and conscience	-.06	Non-significant <i>r</i> s, but mostly in the indicated direction, according to the text
Sears et al. (1957)	5 B, G (M)	Kindergarten sample	160	Extent of PP, combining severity and frequency (severe PP)	Conscience	-.41	Correlations of high vs. low use of each disciplinary tactic with conscience, similar to Gershoff
				Reasoning	Conscience	.37	Same
				Privilege removal	Conscience	-.14	Same
				Isolation	Conscience	.00	Non-significant <i>r</i>
				Love withdrawal	Conscience	.18	Correlations of high vs. low love withdrawal & conscience
Yarrow et al. (1968)	4 B, G (M)	Nursery school sample	86	Reported use of physical punishment from vignettes about extreme disobedience (conditional PP)	Conscience (maternal report)	-.02	Correlation. This outcome not in Gershoff
				Use of reasoning from vignettes	Conscience	.22	Correlation
				Use of scolding from vignettes	Conscience	-.18	Same
				Use of privilege removal from vignettes	Conscience	.22	Same
				Use of isolation from vignettes	Conscience	-.30	Same
				Use of diverting from vignettes	Conscience	.32	Same
				Use of love withdrawal from vignettes	Conscience	-.04	Same

Table 1. Continued

Study	Age, Gender (Parent) ^a	Sample	N	Discipline tactic ^b	Outcome	Effect size ^c (d)	Basis of effect size & discrepancies from Gershoff (2002)
<i>Prosocial behavior</i>							
Uncontrolled longitudinal studies							
Zahn-Waxler, Radke-Yarrow, and King (1979)	15–24 months B, G (M)	Volunteers	16	Proportional use of PP in emotionally charged mother-child interactions (predominant PP) Explanations with affect (proportional use)	Reparations and altruism 4.5 months later	-.58	Average correlations for time-lag data for reparations and for altruism. Gershoff probably used these plus contemporaneous <i>r</i> s
				Positive suggestions (proportional use) Explanations with neutral affect (proportional use)	Reparations and altruism later	.56	Same
				Physical restraint (proportional use) Unexplained verbal prohibitions (proportional use)	Reparations and altruism later on	-.10	Same
				Ignoring (proportional use)	Reparations and altruism later on	.26	Same
				Parental report of spanking and possibly time out before age 6 (customary PP)	Reparations and altruism later on	-.70	Same
				Parental report of withdrawal of privileges and assigning extra duties before age 6	Reparations and altruism later on	-.34	Same
Retrospective studies							
Watson (1989)	0–5 at T1; 17 at T2; B, G (M, F)	National Merit Scholarship finalists & average test-takers	2500	Parental report of spanking and possibly time out before age 6 (customary PP)	Youth-reported altruism	.00	Correlation, using .00 for non-significant <i>r</i> s. Outcome not in Gershoff
				Parental report of withdrawal of privileges and assigning extra duties before age 6	Youth-reported altruism	.00	Non-significant <i>r</i> s
Uncontrolled cross-sectional studies							
Hall (1994)	4–5 B, G (M)	Preschoolers from high income families	41	Frequency of spanking or slapping, from Conflict Tactics Scale (CTS) items (customary PP) Reasoning (% of maximum possible score on three CTS items)	Verbal positives on interpersonal problem solving task	-.18	Correlation. Gershoff used only the <i>r</i> with nonverbal negatives on a similar task, but the equivalent <i>r</i> was not reported for reasoning
					Verbal positives on interpersonal problem solving task	.44	Correlation

Study	Sample	Age	Outcome	Correlation	Notes
<i>Self-esteem</i> Retrospective studies Larzelere, Klein, Schumm, and Alibrando (1989)	Home Economics college students	0-12 at T1, M = 21 at T2; B, G (M, F)	Spanking frequency (customary PP)	-.10	Correlation in full report mentioned in article's footnote. Outcome not in Gershoff due to non-significant <i>r</i> s
		Alternative punishments (time-out, privilege removal, restitution) in all three age groups	Self-esteem	-.08	Correlation
Watson (1989)	National Merit Scholarship finalists & average test-takers	0-5 at T1; 17 at T2; B, G (M, F)	Parental report of spanking and possibly time out before age 6 (customary PP)	-.03	Correlations with three outcomes, using .00 for non-significant <i>r</i> s. Outcome not in Gershoff
		Parental report of withdrawal of privileges and assigning extra duties before age 6	Youth-reported neuroticism, self-acceptance, and sense of well-being	.00	Non-significant <i>r</i> s
Uncontrolled cross-sectional studies Coopersmith (1967)	Fifth-graders selected for consistency or inconsistency on measures of self-esteem	10-12 B (M)	Predominant use of physical punishment rather than next two tactics, when rules are violated (predominant PP)	-.42	Proportions of the high and low self-esteem groups with PP as predominant disciplinary method, compared to the next two, tactics similar to Gershoff
			Predominant use of love withdrawal	Self-esteem	-.54
<i>Competency</i> Uncontrolled longitudinal studies Crowne, Conn, Marlowe, and Edwards (1969)	Kindergarten children	5 at T1, 18 at T2; B, G (M, F)	Predominant use of milder management tactics ("restraint, denial, isolation") more than above two tactics	.86	Proportions of the extreme self-esteem groups using milder tactics predominantly
			Stress discussion and reasoning to obtain compliance and cooperation, rather than force or autocratic means	Self-esteem	.87
			Frequency and severity of spanking by each parent (severe PP)	-.00	Mean of 15 associations for three measures of physical punishment and five unambiguous outcomes. <i>d</i> = .00 if non-significant. Not in Gershoff

Table I. Continued

Study	Age, Gender (Parent) ^a	Sample	N	Discipline tactic ^b	Outcome	Effect size ^c (<i>d</i>)	Basis of effect size & discrepancies from Gershoff (2002)
Retrospective studies Watson (1989)	0-5 at T1; 17 at T2; B, G (M, F)	National Merit Scholarship finalists & average test-takers	2500	Reasoning	Ambitious, realistic aspirations, etc.	-.09	Mean of five associations (only one measure of reasoning)
				Privilege removal	Ambitious, realistic aspirations, etc.	.00	Mean of five associations
				Isolation	Ambitious, realistic aspirations, etc.	.00	Mean of five associations
				Love withdrawal	Ambitious, realistic aspirations, etc.	.00	Mean of five associations
Uncontrolled cross-sectional studies Hall (1994)	4-5 B, G (M)	Preschoolers from high income families	41	Parental report of spanking and possibly time out before age 6 (customary PP)	Score on National Merit Scholarship Test and reported % rank in class	-.11	Mean of two correlations. Outcome not in Gershoff
				Parental report of withdrawal of privileges and assigning extra duties before age 6	Score on National Merit Test and % rank in class	-.19	Mean of two correlations
Uncontrolled cross-sectional studies Hall (1994)	4-5 B, G (M)	Preschoolers from high income families	41	Frequency of spanking or slapping on Conflict Tactics Scale (CTS) item (customary PP)	Number of relevant solutions in interpersonal conflict task	.73	Correlation. Outcome not in Gershoff
				Reasoning (% of maximum possible score on three CTS items)	Number of relevant solutions in interpersonal conflict task	.73	Correlation

^aAge in years unless otherwise indicated; *M* = mean. T1 = Time 1, T2 = Time 2; B = boys, G = girls; (M) = mothers, (F) = fathers, (B) = both.

^bPhysical punishment is categorized as either conditional, customary, severe, or predominant usage.

^cA positive *d* indicates a beneficial association, i.e., that greater use of the disciplinary tactic is associated with preferable child outcomes, e.g., the tactic is associated with greater prosocial behavior or less antisocial behavior. A negative *d* indicates a detrimental association between the tactic and the child outcome.

^dThe mean pre-post gain for a child-determined release (1.295) was subtracted from these *ds* for analyses of effect sizes of physical punishment (e.g., Table II).

about discipline. Seven of the 26 studies in this meta-analysis were included only in the Larzelere (2000) review. They were probably excluded from the Gershoff (2002) review due to having unusual child outcome variables (three studies of substance abuse, need for power, or realistic/ambitious aspirations), falling outside of her search criteria (two studies), being unavailable via interlibrary loan (1), or being part of a study that was already included (1).

Moderating Variables

The following variables were coded to determine whether they accounted for differences in the effect sizes of physical punishment and alternative disciplinary tactics. A study's *design* was coded as either (1) a randomized experiment, (2) a non-randomized study that controlled for initial child misbehavior with statistical controls or within-subject analyses, (3) a time-ordered study (longitudinal, retrospective, or sequential) in which the measure of the disciplinary tactic clearly preceded the child outcome measure (without controlling for initial misbehavior), or (4) a cross-sectional design, in which the referent periods for disciplinary tactics and the child outcome overlapped in time.

Four types of physical punishment were distinguished. *Conditional* spanking was defined as physical punishment that was used primarily to back-up milder disciplinary tactics (e.g., reasoning or timeout), used for defiance, or used in a controlled manner. These definitions of conditional spanking emerged because each type demonstrated more beneficial outcomes (or less detrimental outcomes) than other types of physical punishment in at least one study (e.g., Larzelere, Schneider, Larson, & Pike, 1996; Ritchie, 1999; Straus & Mouradian, 1998). Although the most optimal usage might incorporate all three definitions, no study explicitly incorporated more than one of these definitions in its measure of physical punishment. *Customary* physical punishment was defined as typical parental usage (e.g., usage or frequency), without emphasizing its severity or predominance. It could have included severe physical punishment, but only to the extent typical of ordinary usage by parents. *Overly severe* physical punishment was based on measures that gave extra points for the severity of physical punishment. Examples included "shaking" (Burton, Maccoby, & Allinsmith, 1961), "severe spankings" (Sears, Maccoby, & Levin, 1957), or spanking when "so angry that you 'lost it'"

(Straus & Mouradian, 1998). Finally, *predominant use* of physical punishment included studies investigating predominant disciplinary tactics (e.g., "the primary disciplinary tactic used") or proportional usage (e.g., the proportion of disciplinary incidents for which the parents used physical punishment rather than milder disciplinary tactics).

Outcome variables were grouped into four categories, consisting of compliance; antisocial behavior (including substance use and abuse); conscience or resistance to temptation; and positive behaviors, competencies, or emotions. Several analyses combined antisocial behavior and conscience into a larger category of misbehavior inhibition, to increase statistical power for testing other moderating variables.

Same-source bias was coded when an effect size was based solely on information provided by the same person, as opposed to incorporating distinct sources of information. Two *age* groups distinguished children averaging older or younger than 7 years at the time of the discipline.

Selected Meta-Analytic Details

Most effect sizes were calculated using Johnson's (1989) DSTAT program, following Gershoff (2002). When a study had multiple relevant statistics, we selected statistics that minimized the methodological problems noted by Baumrind et al. (2002), including the same-source bias and correlational statistics. When a study included several statistics that differed on these characteristics, effect sizes were based on the stronger evidence and also distinguished four types of physical punishment (conditional, customary, severe, and predominant usage). In four studies, the best estimate of the effect size controlled statistically for one or more other variables, such as initial child misbehavior. In those cases, the effect size was based on a standardized regression coefficient or similar statistic, following Glass, McGaw, and Smith (1981). In three of those four studies, the effect size could not be estimated in standard deviations units of the outcome variable. Therefore, those studies were coded as using a distinct standard deviation unit typical of covariance-corrected coefficients, following Glass et al. (1981). Because the effect sizes from those three studies did not differ significantly from the other 23 studies, the distinction was dropped for the main analyses.

Effect sizes (*ds*) were corrected for an upward bias in small studies, using Hedges' correction

(Lipsey & Wilson, 2001). For the analyses, each effect size was weighted by a function of its sample size and the inverse of its extremity (Lipsey & Wilson, 2001, p. 49; Shadish & Haddock, 1994, p. 268).

The calculations of effect sizes from three sets of studies warrant additional clarification, which is provided in the appendix (the Roberts series of studies, two Larzelere studies, and Ritchie, 1999). The guiding principle was to base effect sizes on equivalent analyses of physical punishment and alternative tactics. In addition, Larzelere et al. (1996) and Ritchie (1999) yielded different effect sizes for conditional and customary use of physical punishment.

Two effect sizes were included from a study when they were relevant for distinct cells in a particular analysis. For example, several studies had effect sizes for conditional spanking and for another category of physical punishment. Those studies then contributed two different effect sizes for those two types of physical punishment. This increased the sample size from 26 studies to 32 relevant effect sizes in the initial analyses of type of physical punishment.

Following Hedges (1994), the Q statistic was based on ANOVA sums of squares to test hypotheses about whether weighted mean effect sizes varied significantly by moderating factors. The Q statistic is distributed as χ^2 under the null hypothesis. Most analyses of moderating variables had missing cells because all combinations of those factors were not represented by at least one study. Consequently, the Q statistic was based on the Type IV sums of squares in the weighted fixed-effects ANOVA. Statistical tests of whether weighted differential means differed from zero used a z statistic, based on Lipsey and Wilson (2001, p. 115).

RESULTS

Effect Sizes of Physical Punishment by Research Design and Physical Punishment Type

The effect sizes of physical punishment on child outcomes varied significantly by type of physical punishment, $Q(3) = 9.80$, $p < .05$, by research design, $Q(3) = 28.25$, $p < .001$, and by their interaction, $Q(5) = 17.82$, $p < .01$. As shown in Table II, weighted mean effect sizes appeared detrimental for severe physical punishment ($d = -.22$) and predominant physical punishment ($d = -.21$), but were near zero for customary and conditional physical punishment ($ds = .06$ and $.05$, respectively). Mean effect sizes were apparently detrimental in studies using

correlational designs ($d = -.22$), approached zero in time-ordered and controlled designs ($ds = .10$ and $-.08$, respectively), and were apparently beneficial in studies employing randomized designs ($d = .80$). The interaction effect was due to the following exceptions to the usual pattern of effect sizes become less detrimental or more beneficial as design quality improved: Customary physical punishment produced its most detrimental effect size in statistically controlled studies rather than in cross-sectional studies, whereas predominant usage yielded its most detrimental effect size in time-ordered designs rather than in cross-sectional studies.

Differential Effect Sizes by Research Design and Physical Punishment Type

The next analysis investigated *differential* effect sizes by research design and type of physical punishment. A differential effect size is the difference between the mean effect size for physical punishment and the mean effect size for alternative disciplinary tactics in the same study using the same methodology. The results showed that neither design nor the Design \times Physical Punishment Type interaction was significant, $Q(3) = 4.37$ and $Q(5) = 8.12$, respectively. Differential effect sizes varied only by the type of physical punishment, $Q(3) = 18.26$, $p < .001$.

Table III shows that differential effect sizes favored physical punishment over alternative tactics when physical punishment was defined as conditional (differential $d = .29$) or customary (differential $d = .14$). (For brevity, differential d will be shortened to d from here on, which is what it would be called to describe differences between treatment and control conditions.) In contrast, differential effect sizes favored alternative tactics over both overly severe ($d = -.07$) and predominant physical punishment ($d = -.33$).

It is instructive at this point to compare the results of the first two analyses. In Table II, the effect sizes associated with physical punishment varied significantly by research design and by the Design \times Physical Punishment Type interaction. In Table III, however, differential effect sizes (d for physical punishment minus d for alternative tactics) did not vary significantly by research design or by its interaction with physical punishment type. These results indicate partial success in reducing confounds associated with correlational evidence by using differential

Table II. Weighted Effect Sizes for Physical Punishment by Research Design and Type of Physical Punishment

Type of physical punishment	Research design				Weighted mean <i>d</i>
	Cross-sectional	Time-ordered	Controlled	Randomized	
Predominant	-.22 (4)	-.58 (1)	.05 (1)	—	-.21 (6)
Overly severe	-.32 (3)	.07 (4)	—	—	-.22 (7)
Customary	.06 (2)	.10 (4)	-.19 (3)	—	.06 (9)
Conditional	-.13 (2)	.30 (2)	.68 (2)	.80 (4)	.05 (10)
Weighted mean <i>d</i>	-.22 (11)	.10 (11)	-.08 (6)	.80 (4)	.00 (32)

Note. *n* of studies in parentheses. A positive effect size (*d*) indicates that higher physical punishment scores are associated with more beneficial or less detrimental child outcomes than are low scores on physical punishment. All mean *d*s are weighted by Lipsey & Wilson's equation (2001, p. 49).

effect sizes of alternative tactics from the same study.

To obtain sufficient statistical power to investigate hypothesized moderators of the relative effectiveness of physical punishment and alternative tactics, the following analyses drop research design as a factor because it was not a significant predictor. Likewise, predominant usage and severe physical punishment were combined into one category in subsequent analyses.

Differential Effect Sizes by Outcome and Physical Punishment Type

The next set of analyses investigated whether the differential effect sizes of physical punishment vs. alternatives varied by type of child outcome. The results indicated that outcome type, physical punishment type, and their interaction were significantly related to differential effect sizes: physical punishment type, $Q(2) = 41.15$; outcome type, $Q(2) = 21.15$; interaction, $Q(3) = 16.95$, all $ps < .001$. As expected, conditional spanking showed a more positive differential effect size ($d = .29$) than customary physical punishment ($d = .14$), which, in turn, pro-

duced a more positive differential effect size than severe/predominant physical punishment ($d = -.12$; see Tables IV–VI). Unexpectedly, effect sizes more strongly favored physical punishment for misbehavior inhibition (antisocial behavior and conscience; $d = .12$) than for either compliance ($d = .00$) or positive behavior and affect ($d = .01$). The interaction effect reflected the fact that the differential effect sizes for compliance varied by type of physical punishment much more than for other outcomes. Both the largest negative and the largest positive differential effect sizes occurred for compliance. Severe/predominant physical punishment compared less favorably with alternatives for compliance than for any other outcome, whereas conditional spanking compared more favorably with alternative tactics for compliance than for any other outcome. For most outcomes, differential effect sizes were positive for conditional spanking, approached zero for customary physical punishment, and were negative for severe or predominant usage.

Two of the weighted means listed above depended heavily upon results from the largest study, a retrospective survey of substance abuse in 5044 military personnel (Tennant, Detels, & Clark, 1975). Because of its unusually large sample size, this study

Table III. Weighted Differential Effect Sizes (Physical Punishment Minus Alternative Tactics) by Research Design and Type of Physical Punishment

Type of physical punishment	Research design				Weighted mean <i>d</i>
	Cross-sectional	Time-ordered	Controlled	Randomized	
Predominant	-.37 (4)	-.78 (1)	.22 (1)	—	-.33 (6)
Overly severe	-.11 (3)	.05 (4)	—	—	-.07 (7)
Customary	-.10 (2)	.18 (4)	-.06 (3)	—	.14 (9)
Conditional	.22 (2)	.44 (2)	.59 (2)	.34 (4)	.29 (10)
Weighted mean <i>d</i>	-.03 (11)	.17 (11)	.02 (6)	.34 (4)	.11 (32)

Note. *n* of studies in parentheses. A positive effect size (*d*) indicates that physical punishment is associated with more beneficial or less detrimental child outcomes than are alternative tactics in the same studies. All mean *d*s are weighted by Lipsey & Wilson's equation (2001, p. 49).

Table IV. Effect Sizes of *Conditional^a* Physical Punishment Compared to Alternative Disciplinary Tactics

Alternative disciplinary tactic	Child outcome	N studies	N children	Mean effect size (<i>d</i>)		Mean difference in effect sizes
				Alternative disciplinary tactic	Conditional ^a physical punishment	
Compliance						
Reasoning	Noncompliance	3 ¹⁻³	152	-.03	.55	.59*** ^b
Verbal prohibition	Immediate compliance	1 ¹	24	-.15	.02	.17
Threats or verbal power assertion	Stop defiance	1 ³	90	.08	.97	.89***
Privilege removal or time out	Subsequent compliance	1 ²	38	.02	.01	-.01
Privilege removal	Stop defiance	1 ³	90	-.33	.97	1.30***
Time out	Stop defiance	1 ³	90	.60	.97	.37
Barrier (room time out)	Compliance to commands & time out	3 ⁴⁻⁶	52	1.04	.84	-.20
Reasoning plus nonphysical punishment	Subsequent compliance	1 ²	38	-.08	.01	.09
Ignoring	Stop defiance	1 ³	90	.02	.97	.95***
Love withdrawal	Immediate compliance	1 ¹	24	.40	.02	-.38
Restraint, physical power assertion	Stop defiance, compliance to commands & TO	2 ^{3,6}	108	.34	.85	.51**
Child release from time out (TO)	Compliance to commands and to time out	2 ^{6,7}	34	.02	.77	.74*
Mean for compliance		7	220	.26	.68	.43*** ^b
Antisocial behavior						
Reasoning	Aggression	2 ^{2,8}	96	-.22	.35	.56**
Scolding	School aggression	1 ⁸	58	-.24	.38	.62**
Privilege removal or time out	Aggression	1 ²	38	.26	.26	-.00
Privilege removal	School aggression	1 ⁸	58	.38	.38	.00
Isolation	School aggression	1 ⁸	58	-.18	.38	.56*
Reasoning plus nonphysical punishment	Aggression	1 ²	38	.63	.26	-.37
Reasoning or nonphysical punishment	Antisocial, impulsivity	1 ⁹	744	-.39	-.14	.25***
Love withdrawal	School aggression	1 ⁸	58	-.24	.38	.62**
Diverting	School aggression	1 ⁸	58	-.47	.38	.85***
Mean for antisocial behavior		3	840	-.35	-.07	.28***

Table IV. Continued

Alternative disciplinary tactic	Child outcome	N studies	N children	Mean effect size (<i>d</i>)		Mean difference in effect sizes
				Alternative disciplinary tactic	Conditional ^a physical punishment	
Conscience						
Reasoning	Conscience	1 ⁸	86	.22	-.02	-.24
Scolding	Conscience	1 ⁸	86	-.18	-.02	.16
Privilege removal	Conscience	1 ⁸	86	.22	-.02	-.24
Isolation	Conscience	1 ⁸	86	-.30	-.02	.28
Love withdrawal	Conscience	1 ⁸	86	-.04	-.02	.02
Diverting	Conscience	1 ⁸	86	.32	-.02	-.34
Mean for conscience		1	86	.04	-.02	-.06
Grand mean		9	1050	-.23	.06	.29***

Note. In the last three columns, effect sizes that are positive indicate that more beneficial outcomes are associated with greater use of (a) alternative disciplinary tactics or (b) conditional physical punishment, or that (c) more beneficial outcomes are associated with conditional physical punishment than with alternative disciplinary tactics. Effect sizes are based on comparable statistics, minimizing the methodological problems noted by Baumrind et al. (2002) whenever possible. Means are weighted by sample size and by effect size extremity according to Lipsey and Wilson's equation (2001, p. 49), with each study contributing one mean of its relevant effect sizes. Some differential effect sizes are not an exact difference of the tabled entries due to rounding.

Studies cited: ¹Chapman and Zahn-Waxler (1982), ²Larzelere et al. (1996), ³Ritchie (1999), ⁴Day and Roberts (1983), ⁵Roberts (1988), ⁶Roberts and Powers (1990), ⁷Bean and Roberts (1981), ⁸Yarrow et al. (1968), ⁹Straus and Mouradian (1998).

^aEither (1) nonabusive backup for milder disciplinary tactics in 2- to 6-year-olds, (2) used specifically for defiance in 3- or 4-year-olds, or (3) used in a controlled manner (not out of control due to anger) with 2-14-year-olds.

^bSignificant heterogeneity of the effect sizes contributing to this mean, *Q* statistic, $p < .05$.

* $p < .05$, significantly different from $d = .00$, *z* statistic (only performed in the right-hand column).

** $p < .01$.

*** $p < .001$.

was weighted eight times more than the median-sized study in this meta-analysis. Excluding this study, the weighted mean differential effect size dropped to $d = .00$ for customary physical punishment and to $d = .02$ for misbehavior inhibition (antisocial/conscience). To avoid an overly favorable comparison between customary physical punishment and nonphysical punishment based solely on this one study, we repeated all analyses excluding the Tennant et al. (1975) study and report the results separately whenever this exclusion changed the findings.

Tables IV-VI distinguish between antisocial behavior and conscience, even though they were treated as misbehavior inhibition in the above analyses due to the small number of studies that investigated conscience. Table IV shows that, compared to alternative tactics, conditional spanking was associated with greater reductions in noncompliance, $d = .43$, $z = 3.08$, $p < .01$, and antisocial behavior, $d = .28$, $z = 4.11$, $p < .001$. Conditional spanking did not differ significantly from alternative tactics in promoting the development of conscience, but this was based on only one cross-sectional study.

Table V shows that, compared to alternatives, customary physical punishment was associated with greater reductions in antisocial behavior, but this result depended upon the largest study (Tennant et al., 1975), $d = .17$ with it, $z = 6.56$, $p < .001$; $d = .03$ without it, $z = 0.81$, *ns*. Otherwise, customary physical punishment was not significantly different from alternative tactics in its associations with other outcomes. Table VI shows that, compared to alternatives, severe/predominant physical punishment was associated with less compliance, $d = -.99$, $z = -4.43$, $p < .001$, conscience, $d = -.36$, $z = -4.48$, $p < .001$, positive behavior and affect, $d = -.36$, $z = -2.28$, $p < .05$, and antisocial behavior, $d = .14$, $z = 2.23$, $p < .05$.

Differential Effect Sizes by Outcome, Physical Punishment Type, and Other Predictors

Additional analyses investigated whether the same-source bias, the child's age, or the short- vs. long-term timing of the outcomes influenced differential effect sizes and whether they modified the

Table V. Effect Sizes of *Customary*^a Physical Punishment Compared to Alternative Disciplinary Tactics

Alternative disciplinary tactic	Child outcome	N studies	N children	Mean effect size (<i>d</i>)		Mean difference in effect sizes
				Alternative disciplinary tactic	Customary ^a physical punishment	
Compliance						
Reasoning	Noncompliance	3 ¹⁻³	152	.07	.04	-.03
Verbal prohibition	Immediate compliance	1 ²	24	-.15	.09	.24
Threats or verbal power assertion	Stop noncompliance	1 ³	90	.02	.07	.05
Privilege removal or time out	Subsequent compliance	1 ¹	38	-.02	-.05	-.03
Privilege removal	Stop noncompliance	1 ³	90	.24	.07	-.17
Time out	Stop noncompliance	1 ³	90	.16	.07	-.09
Ignoring	Stop noncompliance	1 ³	90	.33	.07	-.26
Love withdrawal	Immediate compliance	1 ²	24	.40	.09	-.31
Physical power assertion	Stop noncompliance	1 ³	90	.20	.07	-.13
Mean for compliance		3	152	.10	.04	-.06
Antisocial behavior						
Reasoning	Aggression	1 ¹	38	.01	.08	.07
Non-contact punishment	Aggression or substance abuse	2 ^{1,4}	3594	-.07	.23	.31 ^{***c}
Privilege removal	Antisocial behavior or alcohol usage	2 ^{5,6}	3285	-.13	-.10	.03
Sent to room	Antisocial behavior	1 ⁵	785	-.18	-.23	-.05
Mean for antisocial behavior		4	6879	-.10	.06	.17 ^{***bc}
Conscience						
Reasoning	Conscience & compliance	1 ⁷	90	.06	-.04	-.10
Verbal psychological punishment	Conscience, compliance	1 ⁷	90	-.11	-.04	.07
Love withdrawal	Conscience, compliance	1 ⁷	90	-.06	-.04	.02
Mean for conscience		1	90	-.04	-.04	-.00
Mental health						
Privilege removal	Neuroticism, esteem	1 ⁶	2500	.00	-.03	-.03
Reasoning	Positive behavior and affect	1 ⁸	41	.59	.28	-.31

Table V. Continued

Alternative disciplinary tactic	Child outcome	N studies	N children	Mean effect size (<i>d</i>)		Mean difference in effect sizes
				Alternative disciplinary tactic	Customary ^d physical punishment	
Time out, privilege removal, or restitution	Self-esteem	1 ⁹	157	-.08	-.10	-.02
Privilege removal	Prosocial behavior	1 ⁶	2500	.00	.00	.00
Privilege removal	Academic achievement	1 ⁶	2500	-.19	-.11	.08
Mean for positive behavior and affect		3	2698	-.08	-.05	.03
Grand mean		9	7281	-.08	.06	14 ^{***bc}

Note. See Table IV.

Studies cited: ¹Larzelere et al. (1996), ²Chapman and Zahn-Waxler (1982), ³Ritchie (1999), ⁴Tennant et al. (1975), ⁵Larzelere and Smith (2000), ⁶Watson (1989), ⁷Lytton (1977), ⁸Hall (1994), ⁹Larzelere et al. (1989).

^aCustomary physical punishment refers to typical usage by parents, e.g., via measures of usage or frequency without emphasizing severity or predominant usage.

^bSignificant heterogeneity of the effect sizes contributing to this mean, *Q* statistic, $p < .05$.

^cAfter dropping the largest study (Tennant et al., 1975), the differential effect size is not significantly different from $d = .00$ and the remaining effect sizes are homogeneous.

* $p < .05$, significantly different from $d = .00$, *z* statistic (only performed in the right-hand column).

** $p < .01$.

*** $p < .001$.

main conclusions. For each new predictor, the first question was whether it was associated with differential effect sizes at the zero-order level (e.g., correlations). If so, the second question was whether its independent contribution was significant after controlling for outcome and type of physical punishment.

Same-Source Bias

Without controlling for other factors, differential effect sizes favored physical punishment over alternative tactics more when the same source of information was used for parent and child variables than otherwise, $d = .15$ vs. $d = .03$, $Q(1) = 12.42$, $p < .001$. This difference became non-significant, however, after dropping the largest study (Tennant et al., 1975). The next set of analyses controlled for outcome and type of physical punishment in a three-way analysis of variance. Same-source vs. multiple-source data never predicted differential effect sizes significantly in those analyses, either as a main effect or in statistical interactions with outcome or physical punishment type. This was true regardless of whether the largest study was included. These results suggest that differential effect sizes were successful in minimizing the same-source bias.

Age

Age also predicted differential effect sizes significantly by itself. Surprisingly, effect sizes favored physical punishment over alternatives for school-age children ($d = .20$), but not for preschool children, ($d = .00$, $Q(1) = 27.85$, $p < .001$). This age effect disappeared when the largest study was omitted, however (Tennant et al., 1975). After controlling for outcome and type of physical punishment, age was a significant predictor in only one interaction. With Tennant et al. (1975) excluded, the Age \times Outcome \times Physical Punishment Type interaction was significant, $Q(1) = 6.92$, $p < .01$. This interaction was due to differing effects of the Age \times Outcome interaction by type of physical punishment. For conditional and customary physical punishment, age never predicted differential effect sizes either as a main effect or in an interaction. For severe or predominant physical punishment, the Age \times Outcome interaction was significant, $Q(1) = 8.34$, $p < .01$. In general, severe or predominant physical punishment was more detrimental than alternatives for younger than for older children. The major exception to this was that the most detrimental effect of such physical punishment was on self-esteem in older children, based on one study (Coopersmith, 1967).

Table VI. Effect Sizes of *Severe or Predominant*^d Physical Punishment Compared to Alternative Disciplinary Tactics

Alternative disciplinary tactic	Child outcome	N studies	N children	Mean effect size (<i>d</i>)		Mean difference in effect sizes
				Alternative disciplinary tactic	Severe/Predominant ^a physical punishment	
Compliance Reasoning	Noncompliance	1 ¹	90	.44	-.55	-.99***
Antisocial behavior Reasoning	Antisocial or need for power	2 ^{2,3}	116	-.25	.30	.55**
Privilege removal or time out	Antisocial behavior	1 ²	38	.20	.05	-.15
Privilege removal	Aggression or need for power	2 ^{3,4}	238	-.08	.23	.31*
Reasoning plus nonphysical punishment	Antisocial behavior	1 ²	38	.10	.05	-.05
Reasoning or nonphysical punishment	Antisocial, impulsivity	1 ⁵	744	-.39	-.28	.11
Love withdrawal	Aggression or need for power	2 ^{3,4}	238	.08	.23	.15
Mean for antisocial behavior		4	1020	-.29	-.15	.14*
Conscience Reasoning	Conscience or resistance to temptation	2 ^{6,7}	402	.27	-.34	-.61*** ^b
Scolding	Conscience or resistance to temptation	1 ⁷	76	.08	-.09	-.17
Privilege removal	Conscience or resistance to temptation	3 ⁶⁻⁸	506	-.14	-.33	-.19* ^b
Isolation	Conscience or resistance to temptation	3 ⁶⁻⁸	506	.01	-.33	-.33***
Love withdrawal	Conscience or resistance to temptation	3 ⁶⁻⁸	506	.22	-.33	-.54***
Love withdrawal and reasoning	Internal moral motivation	1 ⁹	120	.00	-.16	-.16
Mean for conscience		4	626	.07	-.30	-.36***
Positive behavior and affect Reasoning	Prosocial behavior, self-esteem, aspirations	3 ¹⁰⁻¹²	162	.31	-.20	-.51** ^b
Verbal prohibition	Prosocial behavior	1 ¹²	16	-.70	-.58	.12

Table VI. Continued

Alternative disciplinary tactic	Child outcome	N studies	N children	Mean effect size (<i>d</i>)		Mean difference in effect sizes
				Alternative disciplinary tactic	Severe/Predominant ^a physical punishment	
Privilege removal	Aspirations (realistic, yet ambitious)	1 ¹¹	83	.00	.00	.00
Isolation	Aspirations (realistic, yet ambitious)	1 ¹¹	83	.00	.00	.00
Ignoring	Prosocial behavior	1 ¹²	16	-.34	-.58	-.24
Restraint	Prosocial behavior	1 ¹²	16	.26	-.58	-.84
Restraint, denial, or isolation	Self-esteem	1 ¹⁰	63	.86	-.42	-1.29***
Love withdrawal	Self-esteem, aspirations	2 ^{10,11}	146	-.23	-.18	.05
Mean for positive behavior and affect		3	162	.15	-.21	-.36* ^b
Grand mean		12	1898	-.10	-.22	-.12* ^b

Note. See Table IV.

Studies cited: ¹Minton et al. (1971), ²Larzelere et al. (1998), ³McClelland and Pilon (1983), ⁴Sears (1961), ⁵Straus and Mouradian (1998), ⁶Sears et al. (1957), ⁷Burton et al. (1961), ⁸Grinder (1962), ⁹Aronfreed (1961), ¹⁰Coopersmith (1967), ¹¹Crowne et al. (1969), ¹²Zahn-Waxler et al. (1979).

^aSevere physical punishment indicates that the measure included extra points for its severity in addition to its frequency or usage; predominant use indicates either that physical punishment was the major tactic used or that it was assessed with a proportional usage measure.

^bSignificant heterogeneity of the effect sizes contributing to this mean, *Q* statistic, $p < .05$.

* $p < .05$, significantly different from $d = .00$, *z* statistic (only performed within the right-hand column).

** $p < .01$.

*** $p < .001$.

Short-Term vs. Long-Term Outcomes

Whether the outcomes were assessed short-term or long-term was completely confounded with research design. Randomized, cross-sectional, and sequential within-subject designs yielded outcomes ranging from immediately to the next day, whereas other designs investigated outcomes at least two months after the disciplinary tactics were used. By itself, the timing of the outcome was significantly related to differential effect sizes only when the Tennant et al. (1975) study was included in the analyses, $Q(1) = 7.44$, $p < .01$. In that case, differential effect sizes favored physical punishment over alternatives in long-term outcomes (differential $d = .12$), but not in short-term outcomes ($d = .00$). After controlling for outcome and type of physical punishment, differential effect sizes continued to favor physical punishment over alternatives more for long-term outcomes, but only when Tennant et al. was included, $Q(1) = 9.71$, $p < .05$.

Comparisons with Specific Alternative Tactics

The analyses to this point have compared physical punishment to all alternative tactics together. This grouping could obscure distinctions among alternative tactics in how favorably they compare with physical punishment. Accordingly, Tables IV–VI summarize the weighted mean effect sizes for specific alternative tactics, compared to conditional spanking (Table IV), to customary physical punishment (Table V), and to severe or predominant physical punishment (Table VI).

Overall, conditional spanking was associated with better child outcomes than were alternative disciplinary tactics, $d = .29$, $p < .001$ (Table IV), but this applied only to reductions in noncompliance and antisocial behavior. Conditional spanking did not differ from any alternative tactic in its association with conscience, according to the only relevant study (Yarrow et al., 1968; overall $d = -.06$, *ns*). Specific differential effect sizes favored conditional spanking

over 10 of 13 alternative tactics in associations with noncompliance, antisocial behavior, or both. The differential effect sizes ranged from 1.30 to $-.38$, depending upon the specific alternative tactic and the outcome. In no case was an alternative tactic associated with significantly less noncompliance or antisocial behavior than conditional spanking.

Combining data for all outcomes, the weighted mean effect sizes for specific tactics favored only 2 of 13 alternative tactics over conditional spanking, albeit non-significantly so. The barrier-enforced backup for time-out had a differential mean d of $-.20$, $z = -.74$, *ns*. The combination of nonphysical punishment and reasoning⁴ had a differential effect size of $d = -.02$, *ns* (Larzelere et al., 1996). Verbal prohibition was the only other disciplinary tactic that did not differ significantly from conditional spanking in its associations with child outcomes. Other mean differential d s ranged from .17 (love withdrawal,⁵ verbal prohibition) to .95 (ignoring), after combining weighted means across all outcomes in Table IV.

Table V shows that customary physical punishment averaged being more effective than alternative disciplinary tactics overall, $d = .14$, $z = 5.84$, $p < .001$. The mean differential effect size would have been $d = .00$, except for the large retrospective study of substance abuse (Tennant et al., 1975), which had a moderately large differential effect size favoring physical over nonphysical punishment, $d = .31$, $p < .001$. All the significant differential effect sizes for customary physical punishment depended upon the large Tennant et al. (1975) study for its significance. Otherwise, no tactic differed significantly from customary physical punishment in its effect size with any outcome, in either direction.

Alternative tactics tended to be associated with better outcomes than severe or predominant physical punishment, although by a surprisingly small degree, $d = -.12$, $z = -2.49$, $p < .05$. Specific tactics that produced significantly better effect sizes than severe/predominant physical punishment for at least one outcome included reasoning, time-out (termed "isolation"⁶ in some older studies), and love with-

drawal. Privilege removal had a significantly better effect size than severe/predominant physical punishment for enhancing conscience, but a significantly worse effect size for reducing antisocial behavior. The other six disciplinary tactics never differed significantly from severe/predominant physical punishment in their associations with any outcome, although the direction of their differential effect sizes favored five alternatives over severe/predominant physical punishment.

DISCUSSION

The major methodological goal of this meta-analysis was to reduce several confounds in this predominantly correlational literature by analyzing differences in effect sizes between physical punishment and alternative disciplinary tactics. The fact that differential effect sizes did not differ by factors such as research design and the same-source bias suggests that this goal was achieved, at least in part. We then evaluated previous conclusions about physical punishment by investigating whether detrimental child outcomes associated with physical punishment are distinctive of physical punishment itself or whether similar outcomes are also found for alternative disciplinary tactics, when investigated with the same research methods.

Results by Type of Physical Punishment

Whether physical punishment compared favorably or unfavorably with other tactics depended on the type of physical punishment. *Conditional* spanking was more strongly associated with reductions in noncompliance or antisocial behavior than 10 of 13 alternative disciplinary tactics. The mean effect size differences favored conditional spanking over alternatives by amounts between Cohen's (1988) small and medium effect sizes, sizable differences for comparisons between alternative treatments such as these ($d = .43$ for noncompliance and $d = .28$ for antisocial behavior). The effect sizes of *customary* physical punishment were neither worse nor better than any alternative tactic, with the exception that Tennant et al. (1975) found physical punishment to be associated with less substance abuse than non-contact punishment. The differential effect sizes favored alternative tactics only in comparisons with *overly severe* or *predominant* use of physical punishment.

⁴Reasoning is defined as verbal attempts to persuade a child to behave appropriately, e.g. using description of consequences or another type of explanation.

⁵Love withdrawal is defined as temporarily withholding expressions of love and nurturance from the child, used as a discipline category in 11 of the studies, from 1957 to 1983.

⁶Isolation is a term used for a disciplinary tactic in older studies, which seems to approximate time-out better than other disciplinary categories from that era.

Conditional spanking compared favorably with alternatives when defined by one of three criteria. Conditional spanking compared most favorably with alternative tactics when it was defined as a response to defiance in 3- and 4-year-olds (mean $d = .54$: Ritchie, 1999; Yarrow et al., 1968) or as a back-up for time-out in clinically oppositional 2–6-year-olds (mean $d = .32$ from Roberts's four studies). Conditional spanking also compared favorably with alternatives when it was defined as controlled usage, not out of control due to anger (mean $d = .25$: Straus & Mouradian, 1998; see also Turner & Muller, 2004). On a fourth criterion, combining spanking with reasoning, the outcomes of conditional spanking were no better than alternative tactics (mean $d = .02$: Chapman & Zahn-Waxler, 1982; Larzelere et al., 1996). It should also be noted that the studies of conditional spanking involved children between 2 and 6 years old, except for Straus and Mouradian (1998), whose sample was 2–14 years of age.

Physical punishment had effect sizes more detrimental than alternatives only when it was used severely or as the predominant disciplinary tactic. This finding supports the consensus statement from the 1996 scientific consensus conference on physical punishment that "Spanking a child should not be the primary or only response to a misbehavior used by a care giver" (Friedman & Schonberg, 1996a, p. 853). It also extends that conference's eighth consensus statement by showing increased risk of dysfunction from severe physical punishment, regardless of the age of the child.

Next, we consider the implications of these results for specific previous conclusions about physical punishment, followed by implications for other disciplinary tactics.

Variations by Outcomes, Duration, and Alternative Tactics

The results of this meta-analysis contradict some previous conclusions about physical punishment and support other conclusions. First, the findings contradict previous conclusions that the only exception to the detrimental effects of physical punishment is for immediate child compliance (Gershoff, 2002; Straus, 2001). Instead, the effect sizes of conditional spanking compared favorably with alternative tactics for all disruptive behavior problems, including antisocial behavior and defiance. Second, physical punishment competed just as well with alternative tactics for long-

term outcomes as for short-term outcomes. In fact, the results favored physical punishment over alternatives more for long-term outcomes than for short-term outcomes, but only when the largest retrospective study (Tennant et al., 1975) was included in the analyses.

Third, all types of physical punishment were associated with *lower* rates of antisocial behavior than were alternative disciplinary tactics. Conditional spanking produced effect sizes more favorable than alternative tactics for subsequent school aggression in 4-year-olds (Yarrow et al., 1968) and for concurrent antisocial behavior in 2–14-year-olds (Straus & Mouradian, 1998). Customary physical punishment was associated with lower substance abuse than were other tactics (Tennant et al., 1975). Even overly severe or predominant physical punishment predicted less antisocial aggression than did alternative tactics, based on two longitudinal studies (McClelland & Pilon, 1983; Sears, 1961) and one cross-sectional study (Straus & Mouradian, 1998). Four other studies found that physical punishment and alternative tactics did not differ in their associations with antisocial behavior. Thus, if physical punishment increases aggression and antisocial behavior, it does so to the same degree or less than the disciplinary tactics to which it has been directly compared.

Fourth, this meta-analysis failed to detect negative side effects unique to physical punishment (Benjet & Kazdin, 2003). This finding is consistent with three reviews that concluded that the negative side effects of punishment are minor and can easily be avoided by making the punishment contingencies clear and by reinforcing appropriate behavior (Matson & Taras, 1989; Newsom, Favell, & Rincover, 1983; Walters & Grusec, 1977). Indeed, two reviews (Matson & Taras, 1989; Newsom et al., 1983) concluded that positive side effects of punishment were more common than negative side effects. One review (Walters & Grusec, 1977) considered physical aggression to be an established negative side effect of physical punishment. On that point, this meta-analysis found no evidence that physical punishment was more strongly associated with physical aggression than other disciplinary tactics.

Two previous conclusions about physical punishment were partially supported. The first was that physical punishment fails to teach positive alternative behaviors. No form of physical punishment was more strongly associated with the development of conscience or of positive behaviors, emotions, or competencies than were alternative tactics. Yet,

compared with other disciplinary tactics, physical punishment predicted lower positive child outcomes only when it was used severely or predominantly, similar to the results for other outcomes.

Second, the analyses partially supported the conclusion that nonphysical punishments are just as effective as physical punishment. The strongest competitors of conditional spanking for reducing behavior problems included the barrier-enforcement method ("room time-out": Day & Roberts, 1983; Roberts, 1988; Roberts & Powers, 1990), a combination of nonphysical punishment and reasoning (Larzelere et al., 1996), and verbal prohibition (only in Chapman & Zahn-Waxler, 1982). In addition, time-out (termed "isolation" in older studies) was equally effective at inhibiting misbehavior in contrasts with conditional and customary physical punishment combined.

The outcomes of nonphysical punishment were not always equivalent to those of physical punishment, however. The meta-analytic results favored conditional spanking over nonphysical punishments in general for reducing defiance and antisocial behavior, mean $d = .34$, $p < .05$. In addition, customary physical punishment was associated with less substance abuse than was non-contact punishment (Tennant et al., 1975).

To summarize this section, this meta-analysis only partially supported prevailing conclusions about physical punishment. Most of the previous evidence against physical punishment does not appear to be unique to physical punishment. Equivalent analyses produce similar evidence against a range of alternative disciplinary tactics as well. This pattern is what would be expected if systematic biases, such as the intervention selection bias, the same-source bias, and the lumping bias, account for most of the evidence against physical punishment. Two previous conclusions were supported in this meta-analysis. First, physical punishment, like other forms of punishment, does not enhance positive development, but only inhibits inappropriate behavior, such as defiance and antisocial behavior. Second, most types of nonphysical punishment had similar associations with outcomes as did physical punishment, although they had better outcomes only in comparisons with overly severe or predominant physical punishment.

If differential effect sizes provide only partial support for prevailing conclusions about physical punishment, what are the implications for research on parental discipline in general? The next section considers these broader implications.

Implications for Parental Discipline Research

The most controversial aspect of parental discipline is whether any kind of punishment ever has a place in optimal parental discipline. If so, what distinguishes effective from counterproductive uses of punishment? Two distinct empirical traditions have answered these questions differently. Developmental psychologists tend to view all forms of punishment negatively, especially compared with disciplinary reasoning (Bee, 1998; Berger & Thompson, 1995; Bornstein & Lamb, 1988; Etaugh & Rathus, 1995; Grolnick, Deci, & Ryan, 1997; Grusec, 1997; Holden, 1997; Kochanska, Padavich, & Koenig, 1996). In contrast, behavioral parent trainers typically teach nonphysical punishment as a core skill and discourage verbal discipline except to clarify instructions and contingencies (Aronfreed, 1968; Axelrod & Apsche, 1983; National Institutes of Health, 1991; Patterson, 1982; Walters & Grusec, 1977). Most developmental psychologists might predict reasoning to have consistently better outcomes than physical punishment, whereas behavioral parent trainers would expect nonphysical punishment to compare favorably with both reasoning and physical punishment. In this meta-analysis, reasoning and nonphysical punishment not only failed to have better outcomes than conditional and customary physical punishment, but neither one was consistently better than the other in competing with physical punishment across all outcomes.

The fact that neither perspective was entirely confirmed by these results suggests the need to improve both perspectives. The two viewpoints complement each other very well in most respects (Larzelere, 2001; Larzelere et al., 1996), and their complementary strengths suggest some constructive directions for future conceptualization and research. The developmental psychology perspective tends to be stronger on ecological validity and has several focused theories with the potential to integrate the two perspectives constructively. Behavioral parent training is stronger on internal validity and on specifying effective implementation of many parental discipline skills.

The respective strengths of both perspectives have corresponding weaknesses. Because most research in development psychology uses correlational methodology, it tends to be weak in internal validity. Such research must rule out plausible alternative explanations to determine whether their apparent evidence against punishment can be accounted for by

child effects, genetic effects, and other non-parental effects. The main deficiency in behavioral parent training is that it does not explain how parent–child interaction can best move beyond clarifying and enforcing contingencies. The correlational evidence in developmental psychology shows that predominantly verbal discipline and well-behaved children tend to go together, even though such correlations may be inadequate for confirming the bidirectional causal effects that lead to that happy conclusion. Behavioral parent training may show how parents can regain reasonably cooperative interactions with oppositional children, but the field has not been as strong at identifying subsequent links in the causal chain toward optimal verbal parent–child interaction, especially of the type desired as children grow toward adulthood.

We suggest that the strengths of both perspectives could be integrated to advance our understanding of parental discipline. This integrated model must be able to distinguish effective from counterproductive disciplinary tactics on the basis of a number of factors in addition to the type of disciplinary tactic. These additional factors might include the parent–child context, characteristics of the child and the situation, and the discriminations made by parents as they make ongoing discipline choices.

Integrative Theories

Several focused theories in developmental psychology suggest complementary roles for disciplinary reasoning and punishment, but their integrative potential has not been fully exploited (Grusec, 1997). According to Baumrind's (1973, 1991) parenting styles, authoritative parenting produces optimal child outcomes with a combination of firm control, nurturance, and good communication. Hoffman's (1977) information processing theory emphasizes disciplinary reasoning combined with an intermediate level of disciplinary firmness. Bell and Harper (1977) provided evidence that parental discipline operates like a control system, in which parents use corrective discipline to get children to return to an acceptable range of behavior. Larzelere's (2001) conditional sequence model applied this control system model to a sequential choice of disciplinary tactics, wherein milder disciplinary tactics are preferred, but are backed up with increasingly forceful tactics when child noncompliance persists beyond what is acceptable to the parent.

Such integrative views of reasoning and punishment can account for the otherwise puzzling results in this meta-analysis and in an article by Grusec and Goodnow (1994). The significance of the Grusec and Goodnow article is that they documented the inconsistent empirical evidence for the presumed superiority of disciplinary reasoning over punishment. Reasoning is associated with better outcomes than punishment in middle-class families, but rarely in working-class families, preschoolers, boys, or temperamentally difficult children. The latter samples all have a higher proportion of children who challenge disciplinary limits, thus eliciting more forceful enforcement tactics. This pattern of results is consistent with the view that disruptive children require more frequent use of punishment to support the ultimate goal of maintaining appropriate cooperation with verbal discipline.

Three findings of this meta-analysis are consistent with these integrative theories. First, physical punishment was associated with better outcomes than most alternative tactics only when it was limited to controlled spanking for defiant responses to milder disciplinary tactics. This implementation of spanking enforces cooperation with milder disciplinary tactics, thereby supporting the goal of appropriate cooperation using milder verbal tactics. Second, physical punishment was associated with better outcomes than alternatives only for reductions in noncompliance and antisocial behavior. Compared to alternatives, no type of physical punishment was associated with positive outcomes, such as conscience development and positive behaviors and feelings.

Third, the relative effect sizes of reasoning and nonphysical punishment suggest they may play complementary roles, consistent with these integrative theories. When reasoning and nonphysical punishment were both compared with physical punishment (Tables IV–VI), reasoning was more effective than nonphysical punishment for enhancing positive child characteristics, but nonphysical punishment was better for inhibiting misbehavior.

Effective Tactic Implementation

Along with integrating theories, future research is needed to investigate how parents can use specific disciplinary skills more effectively. Helping parents to skillfully encourage appropriate behavior, prevent discipline problems, and respond to them

with effective verbal correction should reduce the need for punishment of any kind. Further, more skillful use of nonphysical punishment should reduce the need for physical punishment. The most impressive reductions in physical punishment have occurred after parents were trained in a range of disciplinary skills, including effective use of a last-resort tactic, whether conditional spanking (Eyberg, 1993; Roberts, 1984) or alternative ways to enforce compliance with time out (McNeil, Clemens-Mowrer, Gurwitsch, & Funderburk, 1994; Webster-Stratton, 1990). The frequency of physical punishment decreased far more from these interventions than from laws or recommendations against spanking (Ispa & Halgunseth, 2004; Statistics Sweden, 1996). Therefore, training parents in a range of disciplinary skills may be more effective for both child and parenting outcomes than merely prohibiting traditional last-resort tactics.

Variations in Effectiveness by Situation

Finally, understanding parental discipline would be enhanced by understanding the discriminations by which parents choose one disciplinary tactic over another. Effective parents do not rely on the same disciplinary tactics all the time, but match them to the situation. In Ritchie's (1999) study, mothers of 3-year-olds were most likely to select spanking and time out for defiance, privilege removal for simple refusals or passive noncompliance, and ignoring for passive noncompliance or whining (Larzelere, 2002). In each case, the preferred tactic was optimally effective for terminating that particular form of noncompliance. These findings suggest that mothers choose disciplinary tactics based on finer discriminations than those represented in theories about parental discipline. Perhaps the immediate effectiveness of some of these tactics is outweighed by detrimental long-term effects. This meta-analysis, however, indicates that distinctive long-term effects of physical punishment are similar to its short-term effects.

In any case, parents generally modify their disciplinary tactics based on the perceived effects, but using a much quicker feedback loop than reflected in most research studies. Ritchie (1999) showed that mothers modified their disciplinary tactics when noncompliance persisted within a single disciplinary episode, changing somewhat from verbal tactics to disciplinary punishment. Roberts and Powers (1990) demonstrated the effectiveness of modifying a last-

resort tactic with clinically defiant children. If the children did not comply with time out after six repetitions of the initial back-up tactic, they switched to an alternative back-up tactic, changing from spanking to the barrier-enforcement, or from any other back-up to spanking. In each case, compliance to time out was achieved with the second back-up tactic. The importance for parents to adjust their disciplinary response to the situation was emphasized in the following quote from Grusec and Goodnow (1994):

Hoffman (1970) [observed] that different situations seemed to 'pull' a particular type of discipline from the parent and that this variation of discipline technique by the situation was particularly the case among mothers of children who had a strong moral orientation. The minimal implication is that one may need to look at the nature of the misdeed, and at the connection between misdeed and disciplinary technique, as a part of the answer to how and when differential effectiveness occurs. The large-scale implication is that explanations should be directed toward accounting for why flexibility is effective, rather than being directed only toward the differential effectiveness of the methods themselves (p. 7).

In summary, research on parental discipline generally needs to account for the inconsistency of the evidence both for the superiority of recommended disciplinary tactics and for the inferiority of disfavored disciplinary tactics. Consistent with Grusec and Goodnow's points (1994), future research needs to account for child effects, genetic effects, and ongoing bidirectional influences between children and parents; it needs to make finer discriminations rather than lumping broad categories of disciplinary tactics together; and it needs to account for how parents vary their disciplinary tactics according to the situation, the type of misbehavior, the child's initial response, and other factors. Otherwise, simplistic methods will continue to yield simplistic conclusions that fail to do justice to the complexity of parental discipline.

Limitations

The quality of any meta-analysis depends, of course, on the quality and quantity of the qualifying studies. Indeed, eighteen of the 26 studies in this meta-analysis relied on weak methodologies, providing only zero-order correlations from cross-sectional, retrospective, or longitudinal designs. The other eight studies included four randomized studies,

two longitudinal studies with statistical controls, and two that used within-subject analyses. The randomized studies had the smallest sample sizes, with only eight or nine children per disciplinary tactic. Weighting studies by sample size tended to minimize the influence of these randomized studies relative to the larger, but poorly controlled studies. Moreover, even though this meta-analysis reduced some biases by analyzing differences between the effect sizes of physical punishment and alternative disciplinary tactics, it could not eliminate systematic biases entirely.

Altogether, only 26 studies were found that investigated physical punishment and an alternative tactic in children under the age of 13. Consequently, the mean effect sizes in some cases were based on only one or two studies. The number of studies was insufficient to analyze all relevant predictors simultaneously, necessitating the step-by-step consideration of hypothesized moderators of differences in effect sizes.

Potential Concerns

Given the importance of this topic for children's welfare, it is important to consider the plausibility of alternative explanations for our meta-analytic conclusions. For a predominantly correlational research literature, causal evidence can be supported only to the extent that other plausible explanations have been ruled out (Larzelere et al., 2004; Shadish et al., 2002, chap. 14). The next section considers several alternative interpretations.

Escalation Toward Abuse

One important concern is the possibility of escalation from mild spanking to severe physical punishment. Virtually all professionals oppose overly severe physical punishment, a view consistent with these meta-analytic results. The point of this concern, however, is the inadequacy of evaluating conditional spanking only when it remains within its defined parameters (e.g., used in a controlled manner). Parents may intend to spank in a controlled manner, yet end up inflicting more pain than intended. This type of escalation is not represented in the effect size of conditional spanking once it crosses that line. The escalation issue is mitigated somewhat by the fact that customary physical punishment never compared unfavorably with any alternative tactic, because

customary physical punishment includes escalations to the degree that they occur in typical parental discipline.

Nonetheless, better research is needed on escalation processes within parent-child disciplinary interactions. Most accused parents portray physically abusive incidents as emanating from a discipline incident (Kadushin & Martin, 1981). We know little about escalation processes within discipline incidents, however. Gershoff (2002) found a consistently positive association between physical punishment and physical abuse, but that necessarily follows from defining all physical abuse as instances of physical punishment (100% of abusers then used physical punishment compared to a lower percentage of non-abusers). Aside from that correlation, the linkage is based on the domino fallacy (Baumrind, 1983; Damer, 1980), which holds that any step in an undesirable direction (e.g., spanking or buying on credit) is always undesirable because it increases the possibility of its undesirable extreme (abuse or bankruptcy). There are clearly some types of physical punishment that increase the risk of abuse, but it is not clear that all forms of physical punishment increase that risk.

There is evidence that the likelihood of disciplinary escalation is reduced when milder disciplinary tactics become more effective. Several studies have shown that milder disciplinary tactics become more effective after they are consistently enforced with nonphysical punishment and, if necessary, non-abusive spanking (Larzelere, Sather, Schneider, Larson, & Pike, 1998; Roberts & Powers, 1990). Consequently, one way to reduce the risk of escalation toward abuse is to help parents use milder disciplinary tactics more effectively.

Research has yet to show that a ban on spanking reduces the subsequent rate of physical abuse by parents (Larzelere & Johnson, 1999). Along with other possibilities, it may be that a spanking ban eliminates the kind of age-appropriate nonabusive spanking that helps parents enforce milder disciplinary tactics, thereby increasing the risk of escalating disciplinary interactions. More objective evaluations of spanking bans are needed before they become models for universal dissemination.

Controlled Longitudinal Studies

A second potential concern is that the studies with the strongest causal evidence against physical

punishment were not represented in this meta-analysis, including five longitudinal studies that controlled statistically for the initial level of the outcome. Straus (2001) has argued that these studies compel all professionals to oppose all spanking as a disciplinary option for parents. Unfortunately, these studies did not qualify for this meta-analysis because they did not investigate an alternative disciplinary tactic. Larzelere and Smith (2000), however, took advantage of the fact that the longitudinal cohort analyzed by Straus, Sugarman, and Giles-Sims (1997) also included parallel questions about four alternative tactics. When analyzed in the same manner that they analyzed spanking, the four alternative disciplinary tactics also predicted higher subsequent antisocial behavior, significantly so for grounding, marginally for privilege removal and allowance removal, and non-significantly for sending children to their room. In addition, the effects for spanking and all four alternative tactics became non-significant when the measure of initial antisocial behavior was improved from the trichotomous measure used by Straus et al. (1997) to a continuous measure of externalizing behavior problems.

These findings suggest that child effects (i.e., the intervention selection bias) were only partially controlled in Straus et al. (1997), because their covariate distinguished only among zero, low, and high levels of initial antisocial behavior. Epidemiologists have shown that residual confounding remains when a covariate is converted into a dichotomous or trichotomous variable (Rothman & Greenland, 1998). Similar types of inadequate statistical controls led initially to an overly negative evaluation of Head Start (Campbell & Boruch, 1975; Campbell & Erlebacher, 1970).

The adequacy of the other four controlled longitudinal studies cited by Straus (2001) falls well short of Straus et al. (1997). Two investigated slapping (Brezina, 1999) or spanking teenagers (Simons, Lin, & Gordon, 1998), one failed to control for the initial level of misbehavior (Straus & Paschall, 1998), and the other concluded prominently, "For most children, claims that spanking teaches aggression seem unfounded" (Gunnoe & Mariner, 1997, p. 768). The results of these controlled longitudinal studies are mixed at best, with the strongest evidence against customary spanking applying almost as strongly to other disciplinary tactics. Moreover, the strongest evidence disappears with improved controls for initial level of misbehavior. Nonetheless, more studies are needed that control for initial child misbehav-

ior,⁷ preferably ones that compare the outcomes of nonabusive spanking directly with alternative tactics.

Equivalent Effects

A third objection might be based on the fact that two forms of nonphysical punishment yielded effect sizes equivalent to conditional spanking. The availability of an equally effective nonphysical punishment has been one rationale for dispensing with physical punishment altogether (Graziano, Hamblen, & Plante, 1996; Straus, 2001). Equivalence of effects makes a poor rationale for a spanking ban for several reasons. Disciplinary tactics with equivalent effectiveness overall may each show superior effectiveness for some children in some situations. Indeed, the barrier method, the most effective disciplinary tactic in this meta-analysis, was ineffective with some children, and a child-determined release from time-out, a relatively ineffective disciplinary tactic, was effective for some clinically oppositional children (Roberts & Powers, 1990). When one disciplinary tactic is not working, parents would benefit from having a range of effective alternatives to turn to, as shown by Roberts and Powers (1990). Moreover, effect equivalence would not be considered sufficient to ban a prescription medication, unless the differences in negative side effects clearly favored other equally effective medications in almost all applications.

Conditional Spanking vs. Non-Optimal Alternatives

A final anticipated objection might be that the current meta-analysis is biased in favor of conditional spanking, because it was not compared with optimal forms of alternative disciplinary tactics. This criticism would be appropriate, in part, for two definitions of conditional spanking (controlled usage or combined with reasoning), but not for the others (defiance, enforcing time-out). Controlled spanking was compared with alternative disciplinary tactics, regardless of whether they were used in a controlled manner (Straus & Mouradian, 1998). But even out-of-control physical punishment compared favorably

⁷As this article was going to press, Grogan-Kaylor (2004) published a study of customary spanking that used an econometric longitudinal analysis to control for omitted as well as incorporated variables.

with recommended alternatives in that study, although not to the same degree as controlled spanking. As for spanking in combination with reasoning, it was compared with a range of tactics, only one of which was a parallel combination of nonphysical punishment and reasoning (Larzelere et al., 1996).

In the other two definitions of conditional spanking, alternative tactics were compared with spanking under identical conditions, either responding to defiance (Ritchie, 1999; Yarrow et al., 1968) or to non-compliance with time-out (four studies by Roberts and his colleagues). Note that the effect sizes favored conditional spanking the most under the latter two definitions, where the operational definition did not entail a biased comparison with alternative tactics.

CONCLUSION

The results of this meta-analysis indicate that the detrimental child outcomes previously associated with physical punishment are not unique to physical punishment itself, except when it is used severely or predominantly. In fact, conditional spanking was associated with significantly less defiance or antisocial behavior than 10 of 13 alternative disciplinary tactics. The usual ways that parents use physical punishment (“customary”) were never associated with worse outcomes than any alternative tactic. A possible explanation for these results is that systematic methodological biases account for the correlational evidence linking physical punishment to detrimental child outcomes. When the same methods and logic are used for alternative disciplinary tactics, child outcomes appear equally detrimental.

We are not the first to point out the weak and uneven empirical support underlying prevailing conclusions about parental discipline (Grusec & Goodnow, 1994; Harris, 1998). For recommended disciplinary practices to be based on scientific evidence, researchers must account for the inconsistency of empirical support for prevailing conclusions. Integrative theories can account for the inconsistent data, while reconciling disparate views about disciplinary tactics from two distinct empirical literatures.

Before social scientists conclude that the 94% of American mothers who spank their 3- and 4-year-olds are invariably doing harm to their children (Straus & Stewart, 1999), they should have supportive evidence that is distinctive of nonabusive spanking. Psychologists owe it to parents and their

children to base recommendations for sweeping societal changes on the best scientific evidence possible. Without that safeguard, children are at risk for becoming the victims of well intentioned, but premature policy changes.

APPENDIX: SELECTED EFFECT SIZE DETAILS

The appendix first compares the current study’s effect sizes with those estimated by Gershoff (2002). It then provides additional specifics about our effect size estimates from three sets of studies.

Comparisons of Effect Sizes with Gershoff (2002)

The overall mean weighted effect size for physical punishment in the current meta-analysis was $d = .00$, compared to Gershoff’s (2002) $d = -.35$, which indicated somewhat detrimental associations with child outcomes. The seven studies herein that were not included in Gershoff had a weighted mean of $d = .15$, indicating slightly beneficial associations of physical punishment with child outcomes. The overall beneficial association was produced primarily by the largest study, a retrospective analysis of childhood predictors of adult substance abuse (Tennant et al., 1975). Therefore, the Results section indicates when the findings differ according to whether this study is included or omitted. The remaining six studies yielded a weighted mean of $d = -.09$. The 19 studies that overlapped with Gershoff’s meta-analysis yielded a weighted mean effect size of $d = -.23$ in Gershoff, compared to $d = -.10$ according to our revised effect sizes. Both means indicated slightly detrimental associations, but less so than in Gershoff’s overall meta-analysis.

The differences in effect sizes between the two meta-analyses are due to a variety of factors, two of which stand out. First, only 26% (5 of 19) of the Gershoff (2002) studies in this meta-analysis measured overly severe physical punishment, compared to 65% (34 of 52) of studies with aggression-composite outcomes in her meta-analysis (Baumrind et al., 2002). Studies of overly severe physical punishment were less likely to investigate alternative disciplinary tactics, probably because they emphasized the severity of physical punishment as an indicator of dysfunctional parent-child relations. In contrast, studies of customary or conditional physical

punishment were more likely to investigate alternative disciplinary tactics also, thus qualifying for this meta-analysis.

Second, only 3 of the 19 overlapping studies had similar effect sizes in both meta-analyses. Reasons for the discrepancies resulted from our decisions to include non-significant associations in calculating effect size averages (three studies), to include all relevant associations (two studies), to use pre-post gain scores in the Roberts's series of studies (three overlapping studies), to use the best statistics available (longitudinal associations based on multiple sources of information specific to physical punishment rather than other associations in three studies), to use a different outcome variable (one study), and to use statistics that were similar for alternative tactics (one additional study). We could not replicate Gershoff's estimates in three other studies in which her effect sizes indicated more beneficial outcomes of physical punishment than our estimates. Overall, the mean absolute difference in effect size estimates was .76 for the 19 overlapping studies (unweighted mean). In two extreme cases, our effect size was 2.17 more favorable than Gershoff's *d* for Day and Roberts (1983), whereas our effect size was 4.42 less favorable than Gershoff's *d* for Larzelere et al. (1996). Table I indicates how the effect size estimates in the current meta-analysis differ from Gershoff's estimates.

Effect Size Estimates for Three Sets of Studies

The second part of this appendix specifies how effect sizes were estimated for three sets of studies. These were selected because (1) they included the greatest discrepancies between the effect sizes estimates in this meta-analysis compared to Gershoff's (2002) meta-analysis, (2) the estimating methods could not be easily summarized in one phrase in Table I, (3) and the studies are particularly relevant for conditional spanking.

In the series of studies by Roberts and his colleagues, our effect sizes for each tactic used to enforce time-out were based on two outcomes: compliance to parental clean-up commands (also used by Gershoff) and compliance with time-out in the clinic session. Improvement in compliance to parental commands was estimated with a pre-to-post treatment effect size. The child-determined release condition (Bean & Roberts, 1981; Roberts & Powers, 1990) was treated as the control condition. Therefore, its mean effect size was subtracted from the

effect size for each other tactic, to yield that tactic's additional gain in compliance beyond that produced by a child-determined release from time-out. In three of the four studies, tactics could also be compared on post-treatment measures of the child's compliance with time-out. The final effect size for each tactic was the difference between it and the mean effect size of the child-determined release condition. The purpose of these estimation methods were (1) to estimate the effect sizes using similar methods for each tactic, (2) to have the same standards of comparison for each study by Roberts and his colleagues, and (3) to expand the outcome variables to include compliance with time out as well as compliance to parental commands.

Ritchie's (1999) study was particularly relevant for comparing the ability of eight disciplinary tactics to immediately stop several distinct types of noncompliance. Of particular interest for this meta-analysis were comparisons between the conditional probability that each type of noncompliance immediately *preceded* each disciplinary tactic and the conditional probability that it immediately *followed* that same disciplinary tactic. For example, we wanted to compare the probability that defiance occurred immediately before a spanking with the probability that defiance would be occurring immediately after it was used. The conditional probabilities subsequent to each tactic were given by Ritchie (1999). However, the immediately preceding conditional probabilities had to be estimated from information in the article (Larzelere, 2002). These estimates are only approximate and could be off by as much as 28%, based on the fact that the preceding conditional probabilities summed to totals ranging from 83 to 128%. The preceding conditional probabilities were adjusted for the discrepancy of these totals from 100%, but this merely averaged the estimation errors across the six types of noncompliance. Nonetheless, we estimated effect sizes based on immediate changes in the conditional probabilities of two types of noncompliance. The first type of noncompliance was defiance. The effect size for spanking and changes in defiance were categorized as conditional physical punishment. The second type of noncompliance was called milder noncompliance, based on the average effect size for passive and physical noncompliance. The contrast in effect sizes for different tactics varied quite a bit for defiance vs. milder noncompliance.

We based the effect sizes from Larzelere et al. (1996) from deviation delays, because they provided stronger causal evidence than simple delays until a

misbehavior recurrence (see rationale in Larzelere, 1996). Further, each effect size was calculated relative to the effect size for "other" disciplinary responses, which did not include reasoning or any type of punishment, and was therefore treated as the control condition. When effect sizes were averaged across categories (e.g., for both fighting and noncompliance incidents), they were weighted by the number of discipline incidents in each category.

ACKNOWLEDGMENT

This research was supported in part by Grant 1R03HD044679 from the National Institute of Child Health and Human Development. A previous version of this article was presented at the biennial convention of the Society for Research in Child Development, Tampa, FL, April 2003.

REFERENCES

(References marked with an asterisk indicate studies included in the meta-analysis.)

- *Aronfreed, J. (1961). The nature, variety, and social patterning of moral responses to transgression. *Journal of Abnormal and Social Psychology, 63*, 223–241.
- Aronfreed, J. (1968). Aversive control of socialization. In W. J. Arnold (Ed.), *Nebraska Symposium on Motivation* (Vol. 16, pp. 271–320). Lincoln: University of Nebraska Press.
- Axelrod, S., & Apsche, J. (Eds.). (1983). *The effects of punishment on human behavior*. New York: Academic Press.
- Bauman, L. J., & Friedman, S. B. (1998). Corporal punishment. *Pediatric Clinics of North America, 45*, 403–414.
- Baumrind, D. (1973). The development of instrumental competence through socialization. In A. D. Pick (Ed.), *Minnesota Symposia on Child Psychology* (Vol. 7, pp. 3–46). Minneapolis: University of Minnesota Press.
- Baumrind, D. (1983). Specious causal attributions in the social sciences: The reformulated stepping-stone theory of heroin use as exemplar. *Journal of Personality and Social Psychology, 45*, 1289–1298.
- Baumrind, D. (1991). The influence of parenting style on adolescent competence and substance use. *Journal of Early Adolescence, 11*, 56–95.
- Baumrind, D., Larzelere, R. E., & Cowan, P. A. (2002). Ordinary physical punishment: Is it harmful? Comment on Gershoff (2002). *Psychological Bulletin, 128*, 580–589.
- *Bean, A. W., & Roberts, M. W. (1981). The effect of time-out release contingencies on changes in child noncompliance. *Journal of Abnormal Child Psychology, 9*, 95–105.
- Bee, H. (1998). *Lifespan development* (2nd ed.). Reading, MA: Addison-Wesley.
- Bell, R. Q., & Harper, L. V. (1977). *Child effects on adults*. Hillsdale, NJ: Erlbaum.
- Benjet, C., & Kazdin, A. E. (2003). Spanking children: The controversies, findings, and new directions. *Clinical Psychology Review, 23*, 197–224.
- Berger, K. S., & Thompson, R. A. (1995). *The developing person through childhood and adolescence* (4th ed.). New York: Worth.
- Bornstein, M. H., & Lamb, M. E. (Eds.). (1988). *Developmental psychology: An advanced textbook* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Brezina, T. (1999). Teenage violence toward parents as an adaptation to family strain: Evidence from a national survey of male adolescents. *Youth & Society, 30*, 416–444.
- *Burton, R. V., Maccoby, E. E., & Allinsmith, W. (1961). Antecedents of resistance to temptation in four-year-old children. *Child Development, 32*, 689–710.
- Campbell, D. T., & Boruch, R. F. (1975). Making the case for randomized assignment to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations in compensatory education tend to underestimate effects. In C. A. Bennett & A. A. Lumsdaine (Eds.), *Evaluation and experiment: Some critical issues in assessing social programs* (pp. 195–296). New York: Academic Press.
- Campbell, D. T., & Erlebacher, A. E. (1970). How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuth (Ed.), *Disadvantaged child: Vol. 3. Compensatory education: A national debate* (pp. 185–210). New York: Brunner/Mazel.
- *Chapman, M., & Zahn-Waxler, C. (1982). Young children's compliance and noncompliance to parental discipline in a natural setting. *International Journal of Behavioral Development, 5*, 81–94.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- *Coopersmith, S. (1967). *The antecedents of self-esteem*. San Francisco, CA: Freeman.
- *Crowne, D. P., Conn, L. K., Marlowe, D., & Edwards, C. N. (1969). Some developmental antecedents of level of aspiration. *Journal of Personality, 37*, 73–92.
- Damer, T. E. (1980). *Attacking faulty reasoning*. Belmont, CA: Wadsworth.
- *Day, D. E., & Roberts, M. W. (1983). An analysis of the physical punishment component of a parent training program. *Journal of Abnormal Child Psychology, 11*, 141–152.
- EPOCH-Worldwide. (2004). *Legal reform: Corporal punishment of children in the family*. Retrieved on October 27, 2004 from <http://www.stophitting.com/laws/legalReform.php>.
- Etaugh, C., & Rathus, S. A. (1995). *The world of children*. Orlando, FL: Harcourt Brace.
- Eyberg, S. (1993). *The spank back-up in time-out with preschool children*. Unpublished manuscript, University of Florida, Gainesville.
- Eysenck, H. (1993). Letter to the editor: Hitting the right cause. *The Psychologist, 6*, 392.
- Friedman, S. B., & Schonberg, S. K. (1996a). Consensus statements. *Pediatrics, 98*, 852–853.
- Friedman, S. B., & Schonberg, S. K. (1996b). [Personal statement]. *Pediatrics, 98*, 857–858.
- Gershoff, E. T. (2002). Corporal punishment by parents and associated child behaviors and experiences: A meta-analytic and theoretical review. *Psychological Bulletin, 128*, 539–579.
- Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.
- Graziano, A. M., Hamblen, L., & Plante, W. A. (1996). Subabusive violence in child rearing in middle-class American families. *Pediatrics, 98*, 845–848.
- *Grinder, R. E. (1962). Parental childrearing practices, conscience, and resistance to temptation of sixth grade children. *Child Development, 33*, 803–820.
- Grogan-Kaylor, A. (2004). The effect of corporal punishment on antisocial behavior in children. *Social Work Research, 28*, 153–162.

- Grolnick, W. S., Deci, E. L., & Ryan, R. M. (1997). Internalization within the family: The self-determination theory perspective. In J. E. Grusec & L. Kuczynski (Eds.), *Parenting and children's internalization of values* (pp. 135–161). New York: Wiley.
- Grusec, J. E. (1997). A history of research on parenting strategies and children's internalization of values. In J. E. Grusec & L. Kuczynski (Eds.), *Parenting and children's internalization of values* (pp. 3–22). New York: Wiley.
- Grusec, J. E., & Goodnow, J. J. (1994). Impact of parental discipline methods on the child's internalization of values: A reconceptualization of current points of view. *Developmental Psychology, 30*, 4–19.
- Gunnoe, M. L., & Mariner, C. L. (1997). Toward a developmental-contextual model of the effects of parental spanking on children's aggression. *Archives of Pediatrics and Adolescent Medicine, 151*, 768–775.
- *Hall, E. C. (1994). A correlational analysis of parental conflict resolution practices and 4- and 5-year-old children's interpersonal problem-solving skills and verbal abilities in a preschool setting (Doctoral dissertation, University of San Francisco, 1994). *Dissertation Abstracts International, 55*(12A), 3785.
- Harris, J. R. (1998). *The nurture assumption*. New York: Free Press.
- Hedges, L. V. (1994). Fixed effects models. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 285–300). New York: Russell Sage Foundation.
- Hoffman, M. L. (1970). Conscience, personality, and socialization techniques. *Human Development, 13*, 90–126.
- Hoffman, M. L. (1977). Moral internalization: Current theory and research. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 10, pp. 85–133). New York: Academic Press.
- Holden, G. W. (1997). *Parents and the dynamics of child rearing*. Boulder, CO: Westview Press.
- Ispa, J. M., & Halgunseth, L. C. (2004). Talking about corporal punishment: Nine low-income African American mothers' perspectives. *Early Childhood Research Quarterly, 19*, 463–484.
- Johnson, B. T. (1989). *DSTAT: Software for the meta-analytic review of research literatures*. Hillsdale, NJ: Erlbaum.
- Kadushin, A., & Martin, J. A. (1981). *Child abuse: An interactional event*. New York: Columbia University Press.
- Kochanska, G., Padavich, D. L., & Koenig, A. L. (1996). Children's narratives about hypothetical moral dilemmas and objective measures of their conscience: Mutual relations and socialization antecedents. *Child Development, 67*, 1420–1436.
- Larzelere, R. E. (2000). Child outcomes of nonabusive and customary physical punishment by parents: An updated literature review. *Clinical Child and Family Psychology Review, 3*, 199–221.
- Larzelere, R. E. (2001). Combining love and limits in authoritative parenting. In J. C. Westman (Ed.), *Parenthood in America* (pp. 81–89). Madison: University of Wisconsin Press.
- Larzelere, R. E. (2002). *The effectiveness of alternative disciplinary tactics in reducing various types of noncompliance within extended discipline episodes*. Unpublished manuscript, University of Nebraska Medical Center, Omaha.
- Larzelere, R. E. (2004). *Weighted analyses of the Canadian prevalence of physical punishment by children's age from Cycle 1 of the National Longitudinal Survey of Children and Youth*. Unpublished raw data.
- Larzelere, R. E., Baumrind, D., & Polite, K. (1998). Two emerging perspectives of parental spanking from two 1996 conferences. *Archives of Pediatrics and Adolescent Medicine, 152*, 303–305.
- Larzelere, R. E., & Johnson, B. (1999). Evaluation of the effects of Sweden's spanking ban on physical child abuse rates: A literature review. *Psychological Reports, 85*, 381–392.
- *Larzelere, R. E., Klein, M., Schumm, W. R., & Alibrando, S. A., Jr. (1989). Relations of spanking and other parenting characteristics to self-esteem and perceived fairness of parental discipline. *Psychological Reports, 64*, 1140–1142.
- Larzelere, R. E., Kuhn, B. R., & Johnson, B. (2004). The intervention selection bias: An underrecognized confound in intervention research. *Psychological Bulletin, 130*, 289–303.
- *Larzelere, R. E., Sather, P. R., Schneider, W. N., Larson, D. B., & Pike, P. L. (1998). Punishment enhances reasoning's effectiveness as a disciplinary response to toddlers. *Journal of Marriage and the Family, 60*, 388–403.
- *Larzelere, R. E., Schneider, W. N., Larson, D. B., & Pike, P. L. (1996). The effects of discipline responses in delaying toddler misbehavior recurrences. *Child & Family Behavior Therapy, 18*(3), 35–57.
- *Larzelere, R. E., & Smith, G. L. (2000, August). *Controlled longitudinal effects of five disciplinary tactics on antisocial behavior*. Paper presented at the meeting of the American Psychological Association, Washington, DC.
- Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48*, 1181–1209.
- Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.
- *Lytton, H. (1977). Correlates of compliance and the rudiments of conscience in two-year-old boys. *Canadian Journal of Behavioural Science, 9*, 242–257.
- Matson, J. L., & Taras, M. (1989). A 20 year review of punishment and alternative methods to treat problem behaviors of developmentally delayed persons. *Research in Developmental Disabilities, 10*, 85–104.
- *McClelland, D. C., & Pilon, D. A. (1983). Sources of adult motives in patterns of parent behavior in early childhood. *Journal of Personality and Social Psychology, 44*, 564–574.
- McNeil, C. B., Clemens-Mowrer, L., Gurwitch, R. H., & Funderburk, B. W. (1994). Assessment of a new procedure to prevent timeout escape in preschoolers. *Child & Family Behavior Therapy, 16*(3), 27–35.
- *Minton, C., Kagan, J., & Levine, J. A. (1971). Maternal control and obedience in the two-year-old. *Child Development, 42*, 1873–1894.
- National Institutes of Health. (1991). *Treatment of destructive behaviors in persons with developmental disabilities* (NIH Publication No. 91-2410). Bethesda, MD: Author.
- Newsom, C., Favell, J. E., & Rincover, A. (1983). Side effects of punishment. In S. Axelrod & J. Apsche (Eds.), *The effects of punishment on human behavior* (pp. 285–316). New York: Academic Press.
- Patterson, G. R. (1982). *Coercive family process*. Eugene, OR: Castalia Press.
- *Ritchie, K. L. (1999). Maternal behaviors and cognitions during discipline episodes: A comparison of power bouts and single acts of noncompliance. *Developmental Psychology, 35*, 580–589.
- Roberts, M. W. (1984). An attempt to reduce timeout resistance in young children. *Behavior Therapy, 15*, 210–216.
- *Roberts, M. W. (1988). Enforcing chair timeouts with room timeouts. *Behavior Modification, 12*, 353–370.
- *Roberts, M. W., & Powers, S. W. (1990). Adjusting chair timeout enforcement procedures for oppositional children. *Behavior Therapy, 21*, 257–271.
- Rothman, K. J., & Greenland, S. (1998). *Modern epidemiology* (2nd ed.). Philadelphia: Lippincott-Raven.
- *Sears, R. R. (1961). Relation of early socialization experiences to aggression in middle childhood. *Journal of Abnormal and Social Psychology, 63*, 466–492.
- *Sears, R. R., Maccoby, E. E., & Levin, H. (1957). *Patterns of child-rearing*. New York: Harper & Row.

- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.
- Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis* (pp. 261–281). New York: Russell Sage Foundation.
- Simons, R. L., Lin, K.-H., & Gordon, L. C. (1998). Socialization in the family of origin and male dating violence: A prospective study. *Journal of Marriage and the Family*, *60*, 467–478.
- Smith, D. (2002). Journal article reignites debate over corporal punishment. *Monitor on Psychology*, *33*(8), 14.
- Statistics Sweden. (1996). *Spanking and other forms of physical punishment* (Demography, the Family, and Children 1996:1.2). Stockholm: Author.
- Stattin, H., Janson, H., Klackenber-Larsson, I., & Magnusson, D. (1995). Corporal punishment in everyday life: An intergenerational perspective. In J. McCord (Ed.), *Coercion and punishment in long-term perspectives* (pp. 315–347). Cambridge, England: Cambridge University Press.
- Straus, M. A. (2001). *Beating the devil out of them: Corporal punishment in American families and its effects on children* (2nd ed.). New Brunswick, NJ: Transaction.
- *Straus, M. A., & Mouradian, V. E. (1998). Impulsive corporal punishment by mothers and antisocial behavior and impulsiveness of children. *Behavioral Sciences and the Law*, *16*, 353–374.
- Straus, M. A., & Paschall, M. J. (1998, August 1). *Corporal punishment by mothers and child's cognitive development: A longitudinal study*. Paper presented at the 14th World Congress of Sociology, Montreal, Canada.
- Straus, M. A., & Stewart, J. H. (1999). Corporal punishment by American parents: National data on prevalence, chronicity, severity, and duration, in relation to child and family characteristics. *Clinical Child and Family Psychology Review*, *2*, 55–70.
- Straus, M. A., Sugarman, D. B., & Giles-Sims, J. (1997). Spanking by parents and subsequent antisocial behavior of children. *Archives of Pediatrics and Adolescent Medicine*, *151*, 761–767.
- *Tennant, F. S., Jr., Detels, R., & Clark, V. (1975). Some childhood antecedents of drug and alcohol abuse. *American Journal of Epidemiology*, *102*, 377–385.
- Turner, H. A., & Muller, P. A. (2004). Long-term effects of child corporal punishment on depressive symptoms in young adults: Potential moderators and mediators. *Journal of Family Issues*, *25*, 761–782.
- Walters, G. C., & Grusec, J. E. (1977). *Punishment*. San Francisco: Freeman.
- *Watson, D. G. (1989). Parenting styles and child behavior: A study of retrospective reports from parents of 2500 high school students (Doctoral dissertation, State University of New York at Buffalo, 1989). *Dissertation Abstracts International*, *50*(7B), 3181.
- Webster-Stratton, C. (1990). Enhancing the effectiveness of self-administered videotape parent training for families with conduct-problem children. *Journal of Abnormal Child Psychology*, *18*, 479–492.
- *Yarrow, M. R., Campbell, J. D., & Burton, R. V. (1968). *Child rearing: An inquiry into research and methods*. San Francisco: Jossey-Bass.
- *Zahn-Waxler, C., Radke-Yarrow, M., & King, R. (1979). Child rearing and children's prosocial initiations toward victims of distress. *Child Development*, *50*, 319–330.