Validity of Causal Inferences from Passive Longitudinal Analyses of Corrective Interventions: Accounting for Selection and Regression Artifacts

Robert E. Larzelere Emilio Ferrer
Oklahoma State University University of California at Davis
and
Brett R. Kuhn
University of Nebraska Medical Center

Submitted to the *International Journal of Behavioral Development*
July 23, 2007

Note: This is the originally submitted draft of the following published article:

Larzelere, R. E., Ferrer, E., Kuhn, B. R., & Danelia, K. (2010). Differences in causal estimates from longitudinal analyses of residualized versus simple gain scores: Contrasting controls for selection and regression artifacts. *International Journal of Behavioral Development*, *34*(2), 180-189.

Although the published article is greatly improved over this initial draft, this early draft includes important variations in the analyses that had to be cut due to journal page limitations. These additional analyses included (a) prosocial behavior as a child outcome and (b) three positive parenting variables as predictors. The systematic differences due to selection bias and the corresponding regression artifacts were generally replicated across these other variations except that the bias due to selection bias was reversed for the positive parenting variables, and the magnitudes of the coefficients were smaller.

Abstract

Analyses of passive longitudinal data can yield causally relevant evidence only to the extent that plausible alternative explanations are ruled out. This study compared the ability of five types of longitudinal analyses to correct for selection biases confounded with corrective interventions, using a cohort of 1464 4- and 5-year-olds from Canadian NLSCY data. Three lines of evidence indicated that apparent child outcomes of corrective interventions were due to selection biases and accompanying regression artifacts. First, all significant effects of corrective interventions indicated apparently detrimental effects when predicting residualized change scores, but apparently beneficial effects when predicting simple change scores. This was true whether analyzing measured or latent variables. Second, results from temporally reversed analyses were consistent with selection and regression artifacts, not with unidirectional causal effects. Third, the findings were similar for empirically supported interventions (e.g., *Ritalin*, psychotherapy) and for parental interventions considered controversial (spanking, yelling). Typical longitudinal data may suppress the ability to detect causal effects due to lengthy inter-wave intervals and temporal overlap of covariates and causes. When applicable, longitudinal analyses should check for similar artifacts by implementing temporally reversed analyses and by determining whether the substantive results replicate without artifacts biased in their favor.

Keywords: causal inference, longitudinal analyses, power assertion, antisocial behavior, growth curve analysis, structural equation modeling

Validity of Causal Inferences from Passive Longitudinal Analyses of Corrective Interventions: Accounting for Selection and Regression Artifacts

The most important research questions in developmental psychology are inherently causal even when passive longitudinal data provide the best available evidence. There is little consensus about the best methods to strengthen causal conclusions from such data (McKim & Turner, 1997). Major developmental methodologists have stated that causality can be claimed only from randomized experimental manipulations. To produce cumulative scientific progress typical of other disciplines lacking randomized studies (e.g., astronomy), analyses of passive longitudinal data must discriminate between stronger and weaker evidence for causal inferences. Toward that end, this article investigates the ability of five types of longitudinal analyses to correct for the selection bias inherent in corrective interventions, thereby strengthening their causal evidence.

Simulation studies have shown that analyses of residualized change scores are inherently biased against corrective interventions (e.g., Campbell & Boruch, 1975; Magidson, 2000; Shadish, Cook, & Campbell, 2002), whereas analyses of simple gain scores are biased in favor of them (Lambert & Bickman, 2001). Therefore, we selected two types of analyses to predict residualized change scores and two to predict simple gain scores. Valid causal inferences require the covariates in non-equivalent-groups designs to be measured without error (Campbell & Kenny, 1999), which latent variable analyses accomplish better than analyses of fallible measures (Sörbom, 1979). Therefore each type of change is predicted with one analysis of latent variables and one analysis of measured variables. This yielded the following four representative types of analyses: net-effects multiple regression, a cross-lagged panel analysis of latent variables, correlations with subsequent simple change scores, and a latent growth model. A fifth analysis used longitudinal zero-order correlations, as a baseline analysis for comparison.

We apply these analyses to seven corrective interventions for behavior problems in children, ranging from professional interventions with documented efficacy (e.g., *Ritalin*) to parental interventions often considered counterproductive (e.g., spanking, scolding, and yelling). A corrective intervention is an action selected to correct a perceived problem in a recipient (e.g., psychotherapy, remedial education, disciplinary punishment Larzelere, Kuhn, & Johnson, 2004). Corrective interventions are ideal for this methodological study, because they are inherently confounded with selection biases in most naturally occurring data. The basic question is how to separate the longitudinal effects of these corrective interventions from selection biases and regression artifacts, thereby strengthening causal evidence about these corrective interventions. This study thus reflects the view that causal conclusions are justified to the extent that plausible alternative explanations have been ruled out (Platt, 1964; Shadish et al., 2002).

Although valid causal inferences are not needed to describe developmental trajectories or to predict outcomes from risk factors, they are essential for making valid applications, because applications inherently modify a purported cause to enhance an outcome (Foster & Kalil, 2005). Epidemiologists have made a relevant distinction between a non-causal risk factor (i.e., a marker variable) and a causal risk factor (Kraemer, Stice, Kazdin, Offord, & Kupfer, 2001). A non-causal risk factor can be used to predict future events (e.g., a smooth ear lobe predicts lower risk of heart attacks: *Do you have a crease on your earlobe?*, 2005), but successful interventions require the risk factor to have a causal influence on the outcome. (Cosmetic ear-lobe surgery would not reduce a person’s risk for heart attacks.) Correct applications from longitudinal data require accurate discriminations between causal and non-causal risk factors.

Requirements for Causal Inferences

There are three requirements for valid causal inferences – association, temporal sequence, and isolation (Bollen, 1989; Shadish et al., 2002). There must be an association between the purported cause and effect, the cause must precede the effect, and the purported causal effect must be isolated from alternative plausible explanations of the temporal association. Clarifying the temporal sequence is the major advantage of longitudinal designs, but the correct temporal sequence is not sufficient to establish the causality underlying the association. Most corrective interventions are associated with subsequent detrimental outcomes, whether the intervention is parental, psychological, or medical, because of the intervention selection bias (Larzelere et al., 2004). For example, adolescents who receive mental health treatment are 14 times more likely to commit suicide, according to the median result of nine prospective studies, presumably due to a selection bias. A longitudinal association can therefore superficially indicate a large detrimental effect, even if the corrective intervention is effective in reducing that problem.

Even so, many analyses of longitudinal data fail to keep the temporal sequence as clear as this example, because they analyze cross-sectional associations or associations between simultaneous changes (Rutter, Pickles, Murray, & Eaves, 2001). Moreover, the covariates used to control for pre-existing differences often overlap temporally with the purported cause. If so, the covariate measure may be influenced by the short-term effect of the purported cause, which violates an important assumption for valid causal inferences (J. Heckman, Ichimura, Smith, & Todd, 1998; Huitema, 1980). To the extent covariates reflect an effect of the causal variable, the analysis suppresses its effect on any subsequent outcome. Moreover, if the covariate includes short-term effects from the purported cause, it is impossible to disentangle pre-existing differences from those short-term effects.

A related problem is that the timing of the presumed causal lag is often mismatched with the time interval between longitudinal waves. The unique effect of parental actions during the day of the interview is likely to have faded by the next interview two years later, even though those actions may be the only portion that has not already had a short-term effect on the pre-test covariate. Instead, the causal influence of a particular parental action is likely to be maximized in a matter of days, not years.

Isolation of the purported cause from plausible alternative explanations is the most important, but difficult criterion to satisfy for valid causal inferences. Plausible alternative models must be ruled out to have confidence in a particular causal model (Rutter et al., 2001). That is the central principle underlying epidemiological criteria for causal evidence (e.g., Rothman & Greenland, 1998) and Campbell’s threats to internal validity (Shadish et al., 2002). Both literatures emphasize evidence that plausible alternative explanations have been ruled out, which then strengthens a purported causal effect.

Selection Bias

One of the most common threats to internal validity is a selection bias. Because of that bias, Larzelere et al. (2004) showed that corrective interventions tend to be associated with subsequent detrimental outcomes, whether those interventions are medical (e.g., hospitalization), psychological (marital counseling, treatment for suicide risk), educational (Head Start and remedial education), or parental (power assertive discipline, assistance with homework).

For example, the original summer Head Start program was linked to detrimental academic outcomes in a major initial evaluation (Westinghouse Learning Corporation & Ohio University, 1969). A series of reanalyses of those data showed that standard regression analyses with statistical controls are biased against corrective interventions such as Head Start (Campbell & Boruch, 1975; Magidson, 2000). Campbell referred to this as the *under-adjustment bias*, i.e., that typical statistical controls only reduce a confound and do not eliminate the confound entirely. Epidemiologists refer to the same phenomenon as *residual confounding* (Rothman & Greenland, 1998). This bias appears to be largely overlooked in current psychological research.[1]

The under-adjustment bias occurs because two assumptions for making valid causal inferences from regression-based analyses are rarely satisfied (cf. Freedman, 2006). The first assumption is that the selection process is measured validly and comprehensively. An analysis can use such information to correct for a selection bias by modeling the selection process adequately (J. J. Heckman, 1979). Without comprehensive selection-process information, however, valid causal inferences face “insurmountable” confounds (Rubin, 1978). Longitudinal analyses often control statistically for only a covariate proxy of the actual selection process.

The second assumption is that the covariate is measured without error (Campbell & Kenny, 1999). Measurement error in a covariate increases systematic bias, because the confound is then only partially controlled for, even if the covariate is a valid measure of selection.

Simulation Studies

Several simulation studies have demonstrated that typical statistical controls adjust only partially for the selection bias (Magidson, 2000; Shadish et al., 2002). Campbell and his colleagues (Campbell & Boruch, 1975; Campbell & Kenny, 1999) simulated a selection bias with a correlation of .50 between the pre-test and the post-test. Controlling statistically for the pre-test measure reduced the selection bias by only 50%, resulting in artifactual evidence against a corrective intervention, even when the null hypothesis was true.

In contrast, another simulation demonstrated bias *in favor of* a corrective intervention (Lambert & Bickman, 2001). The crucial difference was that Lambert and Bickman’s (2001) analyses predicted simple change scores from pre-test to post-test rather than predicting residualized change scores (i.e., post-test scores controlling for pre-test scores). Predicting simple change scores is biased in favor of a corrective intervention due to regression toward the mean.

Lord’s Paradox

The contrast between predictions of residualized changes vs. simple gain scores was highlighted in Lord’s (1967) famous paradox. Because his hypothetical data showed no mean change in men’s or women’s weight, analyses of simple gain scores concluded there was no gender difference in change, as shown in Figure 1. However, analyses of residualized changes corrected for pre-exisiting initial differences, thereby concluding that men were heavier and women were lighter at the post-test than predicted by their initial scores. The key to the paradox is that the two types of analyses have distinct implicit counterfactuals (discussed next). Note that Lord’s paradox fits the examples featured in this paper by merely changing the outcome variable to antisocial behavior, labeling the corrective intervention group as the one high in initial antisocial behavior, and labeling the comparison group as the one low in initial antisocial behavior.

Implicit Counterfactuals

Analyses of residualized changes vs. simple gain scores are biased in opposite directions because they assume different counterfactuals, i.e., what would have occurred if the independent variable (e.g., the corrective intervention) had been different. The implicit standard of comparison (counterfactual) in analyses of simple change scores is zero change (or, more precisely, no differential change between the intervention and the comparison groups). Such analyses therefore give credit to a corrective intervention for any differential regression toward the mean, even if it would have occurred without the intervention. In contrast, the implicit null hypothesis in analyses of residualized change scores is differential regression toward the grand mean. This sets a higher standard for how much differential improvement must be associated with the corrective intervention before it is considered effective. Regression toward the grand mean often over-controls for pre-existing differences according to Campbell's under-adjustment bias. From classical test theory, initial scores on extreme groups consist of true scores and measurement error, so they will regress toward *distinct* means (of the true score portions), not toward the grand mean.

By comparison, zero-order longitudinal correlations assume a counterfactual of 100% regression to the grand mean. That is, it assumes that the null hypothesis of no effect would be shown by a corrective intervention when subsequent outcomes are exactly equivalent to those of the comparison group, regardless of pre-existing differences between the groups. That null hypothesis represents such a high standard that only a perfect corrective intervention could satisfy the null hypothesis, and, even then, would not appear more effective than the null hypothesis. A perfect intervention would improve the treatment group so much that they would subsequently be indistinguishable from a comparison group that did not need the corrective intervention. For example, only a perfect treatment for cancer would eradicate symptoms to the extent that the intervention group's post-treatment cancer symptoms were no different than a group not receiving (or needing) the treatment. Yet the resulting longitudinal correlation of $r = .00$ would often be incorrectly interpreted as showing no effect even though it required a perfect intervention. With such an impossibly high standard for the null hypothesis, any less-than-perfect corrective intervention will appear to have detrimental effects according to zero-order longitudinal correlations. Therefore, such correlations between corrective interventions and subsequent outcomes cannot discriminate between effective vs. counterproductive corrective interventions from passive longitudinal data.

Thus longitudinal correlations and analyses of residualized change scores are both biased against corrective interventions. As a result, few corrective interventions are considered effective by developmental psychology, unless there is a substantial number of randomized outcome studies. Examples include the view that all power assertion by parents is detrimental for children (Grolnick, 2003; Grusec, 1997; Kochanska, Padavich, & Koenig, 1996) and that parents should not assist their children with homework (Chen & Stevensen, 1989; Levin et al., 1997). Both are corrective interventions because they are more likely to be used when parents perceive problems in their children.

Temporally Reversed Analyses

Following Campbell and Kenny's (1999) recommendation, all the analyses were re-run after reversing the temporal order of the waves of data. Regression artifacts operate as strongly in reverse as they do forward in time, whereas unidirectional causal effects can only occur forward in time. Therefore regression artifacts would result in a similar pattern of results in backward analyses, whereas unidirectional causal effects would not. Selection biases should also produce similar results in backward analyses, if the selection bias operates equally at all waves, a reasonable assumption in these data.

This study thus investigates the ability of five types of longitudinal analyses to correct for a selection bias in analyzing the effects of seven corrective interventions for children. If the results are due to selection and regression artifacts, then apparent effects should change sign when predicting residualized changes vs. simple gain scores. Second, the pattern of results should be similar when the waves are temporally reversed in the analyses. Finally, the results should be similar for established vs. questionable corrective interventions (e.g., *Ritalin* vs. spanking). On the other hand, unidirectional causal effects would be supported if the associations with outcomes retain the same sign for residualized changes as for simple gain scores, if the pattern of results changes when the analyses are run backwards temporally, and if the analyses discriminate between scientifically established vs. questionable corrective interventions.

Method

Participants

This study analyzed an age cohort from the Canadian National Longitudinal Survey of Children and Youth (NLSCY), a national sample of children up to 11 years old in 1994-1995. The response rate for the longitudinal sample was 92%, 89%, and 79% for the second, third, and fourth waves, respectively. Smaller provinces were over-represented. Weights were available to recover nationally representative estimates, but they were not used in this study to prevent differential weighting from influencing the comparisons among analyses. The 1,464 children in this study included those who were 4 or 5 years old in the first wave, used a parent as the major interviewee at each of the four waves (91.3% mothers, 8.2% fathers in Wave 1), and had complete data on all relevant variables. One child was selected from families with two or more eligible children. Children were excluded from the study if they were identified as having a heart condition, epilepsy, cerebral palsy, or a mental handicap at any wave or if they were receiving special education services at Wave 1. The final sample included slightly more girls than boys (51.3%).

Measures

Child outcomes. The child outcomes included antisocial behavior, hyperactivity, and prosocial behavior, all based on NLSCY scales. Those scales used parent-reported items similar to items on the Child Behavior Checklist, but adapted from major Canadian surveys, such as the Ontario Child Health Study (Boyle, Offord, Racine, Szatmari, & Sanford, 1993) and the Montreal Longitudinal Study (Tremblay, Pihl, Vitaro, & Dobkin, 1994). The items have three possible responses: Never (or Not True), Sometimes (Somewhat True), and Often (Very True). Statistics Canada (2001/2002) formed the scales from factor analyses of Wave-1 data. We recalculated all the scale scores, replacing up to 20% of missing items with that child's mean *z*–score from valid scale items. Standardized *z*-scored items were based on means and standard deviations for 6- and 7-year-olds in Wave 1, so that differences between waves could be retained.

Antisocial behavior was an equally weighted average of NLSCY scales for physical aggression and property offences, based on 6 items each. Sample items on those two scales included "Physically attacks people" and "Destroys others' things," respectively.[2] A square root transformation was used to reduce the skewness of antisocial behavior. Hyperactivity (8 items) and prosocial behavior (10 items) were NLSCY scales, with sample items of "Can't concentrate for a long time" and "Comforts a crying and upset child," respectively (Statistics Canada, 2001/2002). The original coefficient alphas for the four subscales in Wave 1 was .77 for physical aggression, .64 for property offences, .84 for hyperactivity, and .82 for prosocial behavior (Special Surveys Division, 2002). Coefficient alpha for the 12-item scale of antisocial behavior used in this study was .82 (Wave 1).

Professional interventions. The NLSCY included items about three professional interventions considered relevant to the child outcomes. Two parallel items asked, "In the past year, how many times have you seen or talked on the telephone [with the following professionals about this child's] physical or mental health?": "A psychologist or psychiatrist?" and "Any other person trained to provide treatment . . .[e.g.,] a speech therapist, a social worker?" To reduce skewness, the number of visits was grouped into the following four categories: 0, 1, 2 to 9, and 10 or more. Parents were also asked whether their child takes *Ritalin* "on a regular basis" (yes or no).

Corrective interventions by parents. Four parenting variables were considered corrective interventions. Three of them were based on items indicating how often parents used specified disciplinary tactics when their child "breaks the rules or does things he/she is not supposed to." These discipline tactics included "Take away privileges or put in room," "Use physical punishment," and "Raise your voice, scold or yell." The items used 5 alternative responses, ranging from never (scored 1) through sometimes (3) to always (5). The fourth parental corrective intervention was a 7-item scale called hostile-ineffective parenting by the National Longitudinal Survey of Children and Youth (NLSCY), which had a coefficient alpha of .71 in Wave 1 (Special Surveys Division, 2002). The highest loading item was "Has to discipline repeatedly for the same thing." Its items each had 5 alternative responses, somewhat similar to the disciplinary tactic items. If only one item was missing, its value was estimated from the mean *z*-score of the valid items, using the mean and standard deviation for 6- and 7-year-olds at Wave 2.

Positive parenting. Three measures of positive parenting were used, assuming that they would show a distinctive pattern of results due to being less confounded with selection biases. Two of them were NLSCY parenting scales, called positive interaction (5 items) and consistency (5 items), with sample items (and alphas) of "Talk or play with each other" (α = .81) and child "gets away with things" (α = .66), respectively. If there was one missing item per scale, it was replaced using the above strategy. Their coefficient alphas at Wave 1 were .81 and .66, respectively (Special Surveys Division, 2002). In addition, two disciplinary items were averaged to form disciplinary reasoning, including "Calmly discussing the problem" and "Describe alternative ways of behavior," which correlated .46 with each other in Wave 1, $p < .001$.

Statistical Analyses

The possible influence of the professional and parenting variables on the child outcomes were analyzed with 5 types of analyses. The first used zero-order correlations between a Wave 2 variable and a Wave-3 outcome.

The other four analyses controlled for pre-existing outcome differences in distinct ways. Two predicted residualized change scores; two predicted simple gain scores. Within each pair of analyses, one used measured variables at Waves 2 and 3, whereas the other used latent variables based on all four waves.

The analysis of residualized measured change scores was net-effects regression of the Wave-3 outcome regressed on the Wave-2 intervention, controlling for the outcome score at Wave 2. The analysis of residualized latent change scores was based on Jöreskog's cross-lagged panel analysis of latent variables (Jöreskog, ; Jöreskog & Sörbom, 1979) across all four waves (see Figure 2).

The analysis of simple gain scores in measured variables was the correlation between a Wave-2 intervention and a gain score between Wave-2 and Wave-3 outcomes. The analysis of simple latent gain scores was a quadratic latent growth (multilevel) model (see Figure 3). In addition to a quadratic trajectory of the outcome variable over all four waves, it incorporated a pre-post variable coded I = {0, 0, 1, 1} for the four waves to specify the effects of a Wave-2 corrective intervention from pre- to post-intervention occasions. This is similar to a growth curve model with an additional specification for retest or practice effects (Ferrer, Salthouse, Stewart, & Schwartz, 2004). In this model, the coefficient for the intervention variable (y_{px} in Figure 3) represents the effect of the corrective intervention at Wave 2 on the mean subsequent change in the developmental trajectory, controlling for the overall quadratic trend. Note that this predicts simple gains in the latent growth of the outcome scores rather than residualized latent change scores.

The four strongest analyses were re-run after reversing the temporal order of the four waves of data. This is a test of spuriousness due to selection and regression artifacts recommended by Campbell and Kenny (1999), but not yet applied to complex analyses to our knowledge (D. A. Kenny, personal communication, October 14, 2006).

Results

The analyses investigated the apparent effects of seven corrective interventions and three measures of positive parenting on children’s antisocial behavior, hyperactivity, and prosocial behavior. The pattern of results was clearest for associations of the corrective interventions with antisocial behavior and hyperactivity, which will be summarized first. The pattern of results for positive parenting formed a mirror image of the results for corrective interventions, and the pattern of results for prosocial behavior was a mirror image of the results for problem outcomes. The results for positive parenting and for prosocial behavior were generally smaller in magnitude than the results for corrective interventions and behavior problem outcomes.

Results for corrective interventions and problem outcomes differed more by type of analysis than by type of corrective intervention, which is what would be predicted if the results were mostly due to selection and regression artifacts (the first test). We therefore organize the results by type of analysis, but conclude each section by comparing the mean regression coefficients for established and for controversial interventions.

The second test of pervasive selection and regression artifacts was whether the signs of the relevant coefficients would reverse for residualized change scores compared to simple change scores. The short answer is yes, for analyses of both measured and latent variables. All significant coefficients had signs in one direction for residualized change scores and signs in the opposite substantive direction for simple change scores.

Before summarizing those results, we will briefly summarize the baseline analyses of longitudinal correlations. All seven corrective interventions were correlated with higher levels of antisocial behavior and hyperactivity two years later, *p*s < .001 for 13 tests, $p < .05$ for the remaining test (see the first data column in Table 1). The mean longitudinal correlation of the most controversial interventions (physical punishment and scold/yell) with antisocial behavior or hyperactivity (mean $r = .18$) was a little higher than the equivalent mean longitudinal correlation for professional interventions (e.g., psychotherapy or *Ritalin*; mean $r = .13$).

Analyses of Residualized Changes

The general pattern for predicting residualized changes was that all corrective interventions were associated with higher subsequent antisocial behavior and hyperactivity than predicted by pre-existing scores on those outcomes. Compared to longitudinal correlations, statistical controls for pre-existing outcome differences reduced the magnitude of the apparently detrimental associations, but the relevant coefficients never overcame the under-adjustment bias sufficiently to indicate a significantly beneficial effect.

The results of net-effects regression are shown in the third data column of Table 1. These standardized coefficients were smaller than the zero-order longitudinal correlations, but all corrective interventions continued to predict higher levels of antisocial behavior and hyperactivity, significantly so for 10 of the 14 coefficients. The mean coefficient for regressing the two problem behaviors on physical punishment or scold/yell was nearly identical to the equivalent mean coefficient for psychotherapy or *Ritalin* (mean $\beta = .05$ and .06, respectively).

The cross-lagged panel analysis of latent variables showed the same pattern of results, but with even smaller magnitudes (fourth data column, Table 1). Eleven of 14 standardized coefficients indicated that corrective interventions predicted higher levels of problematic behaviors, three significantly, $p < .05$, and four marginally, $p < .10$. No corrective intervention predicted significant reductions in either problematic behavior. The mean coefficient for physical punishment or scold/yell was similar to the mean coefficient for psychotherapy or *Ritalin* (mean $\beta = .02$ and .03, respectively).

Analyses of Simple Gain Scores

Whereas all significant associations of corrective interventions with residualized outcome changes indicated detrimental effects, all significant associations with subsequent simple gain scores indicated beneficial effects. First consider zero-order correlations between a corrective intervention at Wave 2 and simple outcome gain scores from Wave 2 to Wave 3 (second data column, Table 1). Eleven of the 14 coefficients indicated that corrective interventions were associated with subsequent *reduction*s in antisocial behavior or hyperactivity, four of them significantly, $p < .05$, and two marginally, $p < .10$. The mean correlation between physical punishment or scold/yell and changes in antisocial behavior or hyperactivity was -.05, indicating subsequent reductions in those problematic behaviors. The parallel mean correlation for psychotherapy or *Ritalin* was .01.

Figure 4 shows why analyses of simple change scores yield apparent causal conclusions with the opposite sign as apparent causal influences from correlations or analyses of residualized change scores. This figure shows the levels of antisocial behavior at Waves 2 and 3 that were predicted from high, medium, and low levels of nonphysical punishment at Wave 2. The top line shows the mean change in antisocial behavior predicted by frequent non-physical punishment at Wave 2. The next-to-the-top line shows the level of antisocial behavior at Wave 3 predicted by regression toward the mean, from the graphed mean of antisocial behavior at Wave 2. This regression toward the grand mean is approximately the counterfactual implied by net-effects regression, which controlled statistically for Wave-2 antisocial. (A counterfactual can be thought of as the standard of comparison implied by the null hypothesis.)

The three analyses of measured variables in Table 1 yielded different results because they compared the graphed change in antisocial behavior from Wave 2 to Wave 3 with different counterfactuals (standards of comparison). The Wave-3 level of antisocial behavior predicted by high Wave-2 nonphysical punishment was slightly higher than that predicted by regression toward the grand mean, resulting in a small positive β in longitudinal net effects regression (+β in Figure 4). The correlation of Wave-2 nonphysical punishment with change in antisocial from Wave 2 to Wave 3 compared the same predicted value against the no-change standard, yielding a significant negative value for *r* (-*r* in Figure 4). In contrast, the zero-order longitudinal correlation between Wave-2 nonphysical punishment and Wave-3 antisocial behavior compared the latter value against the Wave-3 grand mean, yielding a significant positive value for *r (+r* in Figure 4).[3]

The last column in Table 1 summarizes the results from the latent growth curve model, in which the corrective intervention at Wave 2 predicted subsequent simple gain scores from Wave 2 to Wave 3 controlling for the overall quadratic developmental trajectory of the problem behavior across all four waves. Ten of fourteen coefficients indicated that the corrective interventions were associated with reductions in problematic behavior, although only five of them were significant, $p < .05$. Corrective interventions never predicted significant increases in problem behaviors in the growth curve models. Scold/yell was associated with significant decreases in both problem behaviors, whereas physical punishment was associated with a nonsignificant decrease in antisocial behavior and no change in hyperactivity. Psychotherapy and *Ritalin* were associated with nonsignificant increases in antisocial behavior, but nonsignificant decreases in hyperactivity.

The growth curve models show simultaneously the selection bias and the association of a Wave-2 intervention with subsequent changes in the outcome. For example, Figure 5 shows both the selection biases and the subsequent changes in antisocial behavior associated with high, medium, and low levels of nonphysical punishment at Wave 2. The trajectories show a selection bias in two respects. First, parents used more nonphysical punishment at Wave 2 for children who had been higher in antisocial behavior at Wave 1 (Figure 3's $\gamma_{0x} = 2.30$, $p < .001$). Second, use of nonphysical punishment at Wave 2 was associated with a less rapid decrease in antisocial behavior than their peers from Wave 1 to Wave 2 ($\gamma_{sx} = 1.14$, $p < .05$).[4] Controlling for selection effects in both prior levels and trends as well as the marginally significant curvilinear trend, $\gamma_{qx} = -.20$, $p < .10$, nonphysical punishment at Wave 2 was associated with significant decreases in antisocial behavior from Wave 2 to Wave 3, $\gamma_{px} = -1.52$, $p < .01$. Figure 5 also suggests that the reduction in antisocial behavior due to high use of nonphysical punishment at Wave 2 was maintained from Wave 3 to Wave 4.

Figure 5 also shows why longitudinal correlations with subsequent levels of antisocial behavior suggested detrimental outcomes of nonphysical punishment, yet analyses of subsequent changes in antisocial behavior suggested beneficial outcomes (greater reductions in antisocial behavior). Greater use of non-physical punishment at Wave 2 is associated with more antisocial behavior at Waves 3 and 4, but the zero-order longitudinal correlations do not take into account the selection bias confounded with use of nonphysical punishment by parents.

Summarizing to this point, the results differed more by type of analysis than by type of corrective intervention. Apparently detrimental outcomes of all corrective interventions were found, but only in analyses of residualized changes (or correlations). Apparently beneficial outcomes occurred only in analyses of simple changes, but only for parental discipline (physical punishment only marginally). No beneficial outcomes were found for any professional intervention. Compared to the 2-wave analyses of measured child outcomes, the 4-wave latent-variable models showed reduced magnitudes of the small effects of all corrective interventions, at least for the cross-lagged panel model. Controversial disciplinary interventions by parents (e.g., spanking, scolding or yelling) appeared slightly more detrimental in outcomes than empirically established professional interventions (*Ritalin*, psychotherapy) only in zero-order longitudinal correlations, but not in any of the four stronger analyses.

Analyses of Positive Parenting Variables

To this point, we have summarized the results only for the 7 corrective interventions and the two problem outcomes. The results for the three positive parenting variables (positive interactions, consistency, and reasoning) yielded a similar pattern of results for antisocial behavior and hyperactivity, but in the opposite direction, as expected. Specifically, zero-order correlations indicated that the positive parenting variables were consistently associated with lower levels of problematic behaviors, $p < .05$, but the magnitude of those associations became smaller after controlling for pre-existing outcome differences. After adding those controls, the effects of positive parenting variables were significant in only 2 of the 6 associations in longitudinal net effects (third data column, Table 1) and in 2 of 6 associations in the cross-lagged latent-variable model (fourth data column, Table 1).

Unexpectedly, the direction of association generally reversed when predicting subsequent gains in problematic behavior from the positive parenting variables. When predicting simple gain scores (second data column, Table 1), three of the 6 coefficients predicted increased problematic behaviors, one of them significantly, $p < .05$. In the latent growth curve model (last column, Table 1), four of the six coefficients predicted nonsignificant increases in problematic behavior, one of which approached significance, $p < .10$. Positive parenting never predicted significant or near-significant decreases in subsequent problem behaviors.

Analyses of Prosocial Behavior

Finally, the analyses of prosocial behavior showed the same patterns of results as those summarized above, but with generally smaller magnitudes and reversed signs, as expected. In longitudinal correlations from Wave 2 to Wave 3, positive parenting variables always predicted higher prosocial behavior, $p < .001$, 3 of 4 corrective interventions by parents predicted lower prosocial behavior, $p < .05$, and professional interventions never predicted prosocial behavior. The magnitudes of those effects were generally reduced with longitudinal net effects regression or cross-lagged latent-variable analyses (third and fourth data columns, Table 1). The only significant effect of a professional intervention on prosocial behavior for any of these three types of analyses was that *Ritalin* predicted *lower* prosocial behavior in the cross-lagged latent-variable analysis.

When there were significant effects in predicting simple change scores in prosocial behavior in measured outcomes (second data column, Table 1) or in the latent growth model (last column, Table 1), the direction of effects was reversed so that positive parenting always predicted *decreases* in prosocial behavior, significantly in 4 of 6 coefficients ($p < .05$) and marginally in the other two ($p < .10$). The only significant associations between any corrective intervention and simple change scores in prosocial behavior was that the “hostile/ineffective” scale predicted significant increases in prosocial behavior. Note that all these significant or marginal associations are consistent with regression artifacts.

Temporally Reversed Analyses

The final test of pervasive selection and regression artifacts was to determine whether the results were similar when the analyses were temporally reversed. The results are shown in Table 2, except for longitudinal zero-order correlations. The pattern of results was strikingly similar after reversing the waves of data, exactly what would be predicted from selection biases and regression artifacts. When analyses of residualized changes were implemented in reverse, most results showed that all corrective interventions (e.g., at Wave 3) were associated with higher previous levels of problematic behavior (e.g., at Wave 2), controlling for the concurrent level of the same problematic behavior (e.g., at Wave 3). In contrast, when predicting reversed “gains” in problem behavior, corrective interventions predicted reductions in problematic behavior backwards in time. Indeed, more significant coefficients were consistent with selection and regression artifacts in the temporally reversed analyses than in the original analyses (44 vs. 38 coefficients at $p < .05$).

Table 3 summarizes all the results of the four most causally informative analyses. Of the significant results, 83 are consistent with selection and regression artifacts, whereas only one is inconsistent with them. The one exception was that *Ritalin* predicted higher prosocial behavior previously, even after controlling for the later level of prosocial behavior in the latent cross-lagged panel analysis. Thus the only apparently causal effect that overcame a selection/regression bias suggests that prosocial behavior at any wave predicts a significantly increased likelihood of being on *Ritalin* at the next wave, a finding easily dismissed as a likely Type I error.

Discussion

This study compared five types of longitudinal analyses in estimating the apparent effects of seven corrective interventions and three forms of positive parenting on three child outcomes, focusing particularly on their ability to correct for selection and regression artifacts. The results were consistent with selection and regression artifacts in three ways. First, corrective interventions predicted significant outcomes only in the direction consistent with selection and regression artifacts. When predicting subsequent scores (zero-order correlations) or residualized changes, all significant associations were in the direction of the selection bias, which is generally interpreted as detrimental outcomes for corrective interventions and beneficial outcomes for positive parenting. When predicting simple change scores, however, all significant associations indicated beneficial outcomes for corrective interventions and detrimental outcomes for positive parenting, consistent with regression artifacts. Second, the same pattern of results replicated when run in reverse temporal order, consistent with selection and regression artifacts. Third, the analyses failed to discriminate between empirically supported corrective interventions (e.g., *Ritalin*) and those considered controversial if not counterproductive (e.g, physical punishment, scolding/yelling, “hostile-ineffective” parenting).

Figures 4 and 5 help explain why the apparent effects on simple change scores are the opposite from the apparent effects from analyses of residualized change scores. Both figures show that high nonphysical punishment at Wave 2 is associated with a steeper subsequent decline in antisocial behavior than is low nonphysical punishment, consistent with the apparently beneficial effects of nonphysical punishment on antisocial behavior changes according to the two analyses of simple changes. The contrasting results are illustrated by the fact that high nonphysical punishment at Wave 2 is nonetheless associated with higher *levels* of antisocial behavior at subsequent waves, reflected in apparently detrimental zero-order correlations. Figure 4 compares the observed mean changes in antisocial behavior associated with high nonphysical punishment with the regression toward the grand mean estimated from Wave-2 antisocial behavior. Although high nonphysical punishment at Wave 2 is associated with a greater decrease in antisocial behavior subsequently, the decrease is less than the estimated regression toward the grand mean, yielding a regression coefficient in the detrimental direction in net-effects regression, β = .03, n.s. Thus the figures show why the analyses of subsequent levels or residualized change scores yield conclusions opposite from the analyses of simple gain scores, using identical data.

One advantage of Figures 4 and 5 is that they show the selection bias rather than hiding it. They show that pre-existing differences in antisocial behavior are associated with nonphysical punishment at Wave 2, both concurrently at Wave 2 and historically from Wave 1. The growth trends in Figure 5 also show that differential trends in antisocial behavior between Waves 1 and 2 are associated with nonphysical punishment at Wave 2.

The contrasting results raise the question as to which analysis provides the least biased estimate of the actual causal effects of these corrective interventions. The answer is that the analysis with the correct implicit counterfactual is unbiased. The correct counterfactual reflects how children's antisocial behavior would have changed differently had they received a redced level of the corrective intervention. Unfortunately, there is no way to know that from the data.

Some regression toward the mean seems reasonable, suggesting that analyses of simple changes are biased in favor of corrective interventions. Several considerations, however, suggest that analyses of residualized changes are biased against corrective interventions. In Lord's paradox, males and females constitute two distinct groups, making it unreasonable that they would regression toward the same overall grand mean. Antisocial behavior and the other outcomes in this study do not have two sharply distinct distributions, but it may still be unwarranted to consider such variables as distributions from one homogeneous group, which seems to be an assumption underlying regression toward the mean. Instead, high antisocial groups may be regressing toward a higher mean than well-behaved children are regressing toward. This may be one reason for Campbell's under-adjustment bias. Another factor may be measurement error in the covariate, which violates a crucial assumption for valid causal inferences. The fact that established professional interventions as well as parental interventions had only apparently detrimental outcomes in analyses of residualized change scores suggests that an under-adjustment bias applies across all corrective interventions in these data.

The correct counterfactual is therefore somewhere between zero change as implied by analyses of simple changes and regression toward one grand mean as assumed by analyses of residualized changes. These two types of analyses might therefore constitute upper and lower estimates of the actual causal effect. We do not know how generally that would apply, but it implies that causal evidence should be more impressive when it overcomes both selection biases and regression artifacts. Apparently beneficial causal effects from corrective interventions should be more convincing in analyses of residualized changes, and apparently detrimental causal effects should be more convincing in analyses of simple changes. Although there was no convincing evidence of causal effects in these data, results from some other studies meet that higher standard (e.g., Gunnoe & Mariner, 1997; Pomerantz & Eaton, 2001). It may be worth noting, however, that both of those studies used distinct sources of information for the corrective intervention and the outcome, in contrast to the current study's exclusive use of parental report.

We did not expect to find that selection and regression artifacts also accounted for the apparent effects of positive parenting variables. We anticipated that positive interaction, consistency, and disciplinary reasoning would not be confounded with a selection bias. Moreover, we thought that the beneficial outcomes of positive parenting had more solid causal evidence in the literature. Nonetheless, the apparent effects of positive parenting displayed a pattern consistent with selection and regression artifacts, but in the opposite direction. That is, positive parenting appeared to be confounded with child cooperation and fewer problem behaviors. Their apparently beneficial outcomes in analyses of residualized changes reversed to apparently detrimental outcomes in analyses of simple changes. Moreover, running the analyses backward in time yielded the general pattern predicted by selection and regression artifacts.

These results are generally consistent with the conclusion that the actual detectable unidirectional causal influences of these corrective interventions and positive parenting variable are approximately zero. Because their actual causal effects are so tiny, they had to be inflated by a selection or regression artifact to appear significant, which could seem either beneficial or detrimental, depending upon the direction of the artifact.

This is not to say that the actual causal effects of these corrective interventions are invariably zero. The beneficial effect of many of them have been demonstrated in randomized clinical trials, including some psychotherapies, *Ritalin*, and some forms of disciplinary reasoning and nonphysical and physical punishment. Rather, the point is that the actual causal effects of these corrective interventions, *as typically used*, on these outcomes are not detectably different from zero in this kind of longitudinal study. The near-zero causal effects must be due either to suboptimal implementation or to features of the longitudinal data that suppress the observed causal effects in these analyses. Suboptimal implementation seems inadequate as a complete explanation, because it would imply that ordinary use of all these corrective interventions have the same non-effect, whether they are established professional interventions or controversial parental interventions. This leaves the possibility that causal effects might be suppressed by some features of the longitudinal data.

Although the NLSCY longitudinal data set is optimal in many respects, three typical features of longitudinal data may have suppressed the observed effects of these corrective interventions: measurement adequacy, the time interval between waves, and confounded adjustments for confounds. As is typical of comprehensive longitudinal surveys, a premium was placed on concise measures, which reduces their reliability and validity. Although the reliability of the hyperactivity scale approached that of the corresponding CBCL scale's reliability (α = .82 vs. .84 for CBCL Attention Problems), the reliability of antisocial behavior was slightly lower than the CBCL Aggressive Behavior scale (α = .82 vs. .94). This may be due to the smaller number of items (12 vs. 18) or to the more specific focus of the NLSCY items on physical aggression (Tremblay, 2000). Most of the corrective interventions were measured with only one or two items, which could hinder their reliability.

A greater problem, however, might be the 2-year interval between waves. Consider the case of *Ritalin* and hyperactivity. *Ritalin*'s effect is thought to peak within hours and wane thereafter unless another dose is taken. This causal lag is clearly mismatched with the two-year interval between waves. It may be unreasonable to expect *Ritalin* administered "on a regular basis" recently to produce a causal effect on hyperactivity two years later. Likewise, parents expect to see results from their corrective disciplinary actions within hours or days, effects which may be undetectable two years later.

The third problem is the overlapping referent time periods for the measures of the corrective interventions and pre-existing differences on the outcomes. Parents used similar time referents for whether their child took *Ritalin* on a regular basis and how often their child acted in hyperactive ways. The overlapping weeks or months would incorporate the causal effects of 99% the *Ritalin* dosages during that period. The overlapping referent time periods make it impossible to control for confounds without also controlling for short-term causal effects, which are the strongest effects of *Ritalin*. This violates an assumption for making valid causal inferences, that one should not control for intervening effects of the purported causal variable (J. Heckman et al., 1998; Huitema, 1980)

This explanation suggests that longitudinal data must be collected in a way that permits analyses to pull apart the strongest short-term effect of a corrective intervention from confounds such as a selection bias. Only then can its causal effect on subsequent outcome levels be distinguished from a selection bias. For example, the data collection at each wave might be separated into a series of surveys, with their sequence and referent periods designed to disentangle short-term causal effects from pre-existing differences. At the very least, the questions should be worded with non-overlapping time referents designed to minimize the confound of short-term causal effects with selection biases.

Because such distinctions are rarely made in longitudinal surveys, the usual covariate for pre-existing differences is a combination of a selection bias and as much as 99% of the causal effects, which cannot be disentangled. The remaining 1% of the causal effect would hardly be detectable two years later.

The Case of "Hostile-Ineffective" Parenting

The most discrepant results across analyses were for the apparent effects of "hostile-ineffective" parenting. Any explanation of differential results across analysis types must account for these findings. The key, we believe, is that the items on the "hostile-ineffective" scale are predominantly parent reports about the difficulty of disciplining their child (e.g., "How often are you having problems managing your child in general"). The scale thus measures perceived behavior difficulty, which has been shown to be one of the strongest early longitudinal predictors of subsequent antisocial behavior (Keenan, 2001). Perceived behavior difficulty (aka "hostile-ineffective") may be more strongly confounded with a selection bias than was any corrective intervention, because recognition of a problem is the first step toward selecting a corrective intervention. That confound might explain why hostile-ineffectiveness is the strongest correlate of problem behaviors in cross-sectional correlations, yet turns out to be the strongest predictor of subsequent reductions in the same problem behaviors according to analyses of simple change scores. If a child is showing problematic behavior according to parent reports, it may be better for a parent to recognize that the child is difficult to discipline rather than dismissing such behavior problems as typical (e.g., "boys will be boys"). Consistent with selection and regression artifacts, the apparent effect of perceived behavior difficulty (aka "hostile-ineffective" parenting) on problem behaviors was always significant, but in whatever direction was consistent with selection or regression artifacts.

Limitations

Two limitations of this study should be noted. First, all data were based on parent report. This is advantageous in one sense, in that parental perspectives of child behavior directly influence their choices of corrective interventions. Nonetheless, some associations with child outcomes may have been inflated due to the sole reliance on a single source of information (Yarrow, Campbell, & Burton, 1968). This artifactual inflation could have made the corrective interventions appear more harmful than they actually were, except in analyses of simple change scores. If so, the interventions would appear more effective with outcome measures from an independent source, using analyses of residualized change scores. It remains to be seen whether longitudinal analyses of multi-source data will also produce results consistent with selection and regression artifacts. Using distinct sources of information might be one way to minimize these artifacts.

Second, the data for most corrective interventions were based on one item each. Moreover, the data were skewed for most professional interventions. For example, the percentage of children using *Ritalin* across the four waves was only 0.5% to 2.3% of the sample. Thus, some results could be unstable due to the small sample size of children using *Ritalin* regularly. The percentage of children visiting a psychologist or psychiatrist (0.9% to 6.3%) and the percentage visiting another therapist were a little higher (5.1% to 8.8%). The skewness of these data suggests caution about their estimated standard errors and corresponding statistical tests, but these characteristics should not bias the sign or magnitude of the coefficients for these corrective interventions.

Conclusion

The results from five types of analyses of the same longitudinal data set are consistent with underlying selection biases and regression artifacts instead of unidirectional causal effects in three respects. First, the causal effects switch from apparently detrimental to apparently beneficial outcomes, consistent with selection and regression artifacts. Second, these contradictory apparent effects replicate when the longitudinal analyses are implemented in reverse temporally. Finally, the results fail to discriminate between established corrective interventions by professionals and controversial interventions by parents. The fact that the apparent longitudinal effects depended more on the type of analysis than on the intervention raises important questions about the validity of causal inferences from typical longitudinal analyses, at least when based entirely on parental reports.

These results are important because valid causal inferences are essential for any applications, including those intended to enhance children’s developmental outcomes. The pervasiveness of selection biases might explain why developmental psychology generally opposes all corrective interventions, except for those supported by a substantial number of randomized trials. For example, developmental textbooks usually oppose all power assertive disciplinary tactics except for time out, even though time out has not been supported in any passive longitudinal analyses that we are aware of. As for the most denigrated form of a traditional power assertive tactic, the strongest evidence against customary parental spanking is of the type shown in this study to be plausibly due to selection biases, i.e., effect sizes based on longitudinal correlations (Gershoff, 2002) and coefficients from net-effects longitudinal regression (Straus, 2001). The only statistically controlled longitudinal analyses of customary parental spanking that used distinct sources of information for the outcome variable found that its apparent effects on subsequent aggression varied from beneficial to detrimental depending upon the age, ethnicity, and gender of the child (Gunnoe & Mariner, 1997).

Consistent with the current study, a recent meta-analysis found that the outcomes of customary spanking did not differ from alternative forms of power assertion by parents (Larzelere & Kuhn, 2005). In contrast to that meta-analysis, this study found better outcomes for disciplinary reasoning, but that advantage reversed in analyses of simple change scores, consistent with selection and regression artifacts. Moreover, this study found similar results for established corrective interventions by professionals. This puts the field in the awkward position of empirically supporting an international effort to ban parental spanking based on types of data analyses that provide equally strong evidence against nonphysical punishment, psychological treatment of children, and *Ritalin*.

None of these analyses prove that any type of power assertion by parents is effective, however. Rather they demonstrate that none of these longitudinal analyses can discriminate between effective and counterproductive types of corrective interventions. Recognizing the faulty basis of some current conclusions is necessary, however, to motivate improved research designs and analyses to discriminate between effective and counterproductive interventions.

Researchers must first recognize that the problem of selection and regression artifacts has been consistently under-estimated. Second, they must better understand various ways to minimize such confounds, appreciating what they accomplish as well as their limitations. By recognizing the seriousness of these artifacts and the limitations of currently available statistical adjustments, there will be more support for developing better assessments of selection and regression artifacts, including the temporally reversed analyses in this study (Campbell & Kenny, 1999) and diagnostics based on propensity scores (Dehejia & Wahba, 1999). Improved methods for minimizing selection biases also need to be developed. Promising methods include propensity-based adjustments, such as matching (Rosenbaum, 1995; Schochet & Burghardt, 2007); causal analyses using developmental trajectory analyses (Haviland & Nagin, 2005); and econometric fixed effects models (Wooldridge, 2000). One new possibility implied by the present study is that analyses of residualized vs. simple changes may be biased in opposite directions and may therefore bracket the actual causal effect, at least under some unspecified conditions.

Improved research designs can strengthen causal inferences more than statistical adjustments, especially when those designs pull apart variables that are otherwise confounded (Rutter et al., 2001). Natural experiments and quasi-experiments can enhance causal inferences when randomized experiments are not possible (Shadish et al., 2002). Research also needs to focus more on specific causal mechanisms, a strategy used more in medical research (Rutter, 2003).

The most crucial consideration for making valid causal inferences from passive longitudinal data is to rule out plausible alternative interpretations as much as possible. In the case of corrective interventions by parents, the field has not adequately ruled out the most obvious plausible alternatives, namely selection biases and the regression artifacts that accompany them. Only by minimizing such confounds can we obtain the causal evidence necessary to support applications to enhance child outcomes, especially for those most at-risk.

References

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Boyle, M. H., Offord, D. R., Racine, Y. A., Szatmari, P., & Sanford, M. (1993). Evaluation of the revised Ontario Child Health Study scales. *Journal of Child Psychology & Psychiatry & Allied Disciplines, 34*, 189-213.

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., & Kenny, D. A. (1999). *A primer on regression artifacts*. New York: Guilford Press.

Chen, C., & Stevensen, H. W. (1989). Homework: A cross-cultural examination. *Child Development, 60*, 551-561.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association, 94*, 1053-1062.

Do you have a crease on your earlobe? (2005). Retrieved February 15, 2007, from http://www.mybodylanguage.co.uk/Crease%20earlobe.htm

Ferrer, E., Salthouse, T. A., Stewart, W. F., & Schwartz, B. S. (2004). Modeling age and retest processes in longitudinal studies of cognitive abilities. *Psychology and Aging, 19*, 243-259.

Foster, E. M., & Kalil, A. (2005). Developmental psychology and public policy: Progress and prospects. *Developmental Psychology, 41*, 827-832.

Freedman, D. A. (2006). Statistical models for causation: What inferential leverage do they provide? *Evaluation Review, 30*, 691-713.

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.

Grolnick, W. S. (2003). *The psychology of parental control: How well-meaning parenting backfires*. Mahwah, NJ: Erlbaum.

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.

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.

Haviland, A. M., & Nagin, D. S. (2005). Causal inferences with group based trajectory models. *Psychometrika, 70*, 557-578.

Heckman, J., Ichimura, H., Smith, J., & Todd, P. (1998). Characterizing selection bias using experimental data. *Econometrica, 66*, 1017-1098.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica, 47*, 153-161.

Huitema, B. E. (1980). *The analysis of covariance and alternatives*. New York: Wiley.

Jöreskog, K. G. (1979). Statistical estimation of structural models in longitudinal-developmental investigations. In J. R. Nesselroade & P. B. Baltes (Eds.), *Longitudinal research in the study of behavior and development* (pp. 303-351). New York: Academic Press.

Jöreskog, K. G., & Sörbom, D. (1979). *Advances in factor analysis and structural equation models*. Cambridge, MA: Abt Books.

Keenan, K. (2001). Uncovering preschool precursors to problem behavior. In R. Loeber & D. P. Farrington (Eds.), *Child delinquents: Development, intervention, and service needs* (pp. 117-134). Thousand Oaks, CA: Sage.

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.

Kraemer, H. C., Stice, E., Kazdin, A., Offord, D., & Kupfer, D. (2001). How do risk factors work together? Mediators, moderators, and independent, overlapping, and proxy risk factors. *American Journal of Psychiatry, 158*, 848-856.

Lambert, E. W., & Bickman, L. (2001, March). *Risk adjusted mental health outcomes: Ritual or solution.* Paper presented at the 14th annual research conference, A System of Care for Children's Mental Health: Expanding the Research Base, University of South Florida, Tampa.

Larzelere, R. E., & Kuhn, B. R. (2005). Comparing child outcomes of physical punishment and alternative disciplinary tactics: A meta-analysis. *Clinical Child and Family Psychology Review, 8*, 1-37.

Larzelere, R. E., Kuhn, B. R., & Johnson, B. (2004). The intervention selection bias: An underrecognized confound in intervention research. *Psychological Bulletin, 130*, 289-303.

Levin, I., Levy-Shiff, R., Appelbaum-Peled, T., Katz, I., Komar, M., & Meiran, N. (1997). Antecedents and consequences of maternal involvement in children's homework: A longitudinal analysis. *Journal of Applied Developmental Psychology, 18,* 207-227.

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin, 68,* 304-305.

Magidson, J. (2000). On models used to adjust for preexisting differences. In L. Bickman (Ed.), *Research design: Donald Campbell's legacy* (pp. 181-194). Thousand Oaks, CA: Sage.

McKim, V. R., & Turner, S. P. (Eds.). (1997). *Causality in crisis? Statistical methods and the search for causal knowledge in the social sciences*. Notre Dame, IN: University of Notre Dame Press.

Platt, J. R. (1964). Strong inference. *Science*, *146*, 347-353.

Pomerantz, E. M., & Eaton, M. M. (2001). Maternal intrusive support in the academic context: Transactional socialization processes. *Developmental Psychology, 37,* 174-186.

Rosenbaum, P. R. (1995). *Observational studies*. New York: Springer-Verlag.

Rothman, K. J., & Greenland, S. (1998). *Modern epidemiology* (2nd ed.). Philadelphia: Lippincott-Raven.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, *6*, 34-58.

Rutter, M. (2003). Crucial paths from risk indicator to causal mechanism. In B. B. Lahey, T. E. Moffitt & A. Caspi (Eds.), *Causes of conduct disorder and juvenile delinquency* (pp. 3-26). New York: Guilford.

Rutter, M., Pickles, A., Murray, R., & Eaves, L. (2001). Testing hypotheses on specific environmental causal effects on behavior. *Psychological Bulletin, 127,* 291-324.

Schochet, P. Z., & Burghardt, J. (2007). Using propensity scoring to estimate program-related subgroup impacts in experimental program evaluations. *Evaluation Review, 31*, 95-120.

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

Sörbom, D. (1979). An alternative to the methodology for analysis of covariance. In K. G. Jöreskog & D. Sörbom (Eds.), *Advances in factor analysis and structural equation models* (pp. 219-234). Cambridge, MA: Abt Associates.

Special Surveys Division. (2002). *National Longitudinal Survey of Children and Youth*. Ottawa, ON: Statistics Canada.

Statistics Canada. (2001/2002). *NLSCY Data Users Guide*. Ottawa, ON: Author.

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.

Tremblay, R. E. (2000). The development of aggressive behaviour during childhood: What have we learned in the past century? *International Journal of Behavioral Development, 24*, 129-141.

Tremblay, R. E., Pihl, R. O., Vitaro, F., & Dobkin, P. L. (1994). Predicting early onset male antisocial behavior from preschool behavior. *Archives of General Psychiatry, 51*, 732-739.

Westinghouse Learning Corporation & Ohio University. (1969). *The impact of Head Start: An evaluation of the effects of Head Start on children's cognitive and affective development* (Report presented to the Office of Economic Opportunity Pursuant to Contract of B89-4536, Vols. 1 and 2). Athens, OH: Authors.

Wooldridge, J. M. (2000). *Introductory econometrics*. Stamford, CT: Thomson Learning.

Yarrow, M. R., Campbell, J. D., & Burton, R. V. (1968). *Child rearing: An inquiry into research and methods*. San Francisco: Jossey-Bass.

Author Note

Robert E. Larzelere, Department of Human Development and Family Science, Oklahoma State University. Emilio Ferrer, Department of Psychology, University of California at Davis. Brett R. Kuhn, Psychology Department, Munroe Meyer Institute, University of Nebraska Medical Center.

This research was supported by the National Institute of Child Health and Human Development (R03 HD044679), using data generously made available by Statistics Canada. Thanks are due to Michael Criss, Patricia Cohen, and Michael Rutter for critiquing previous drafts of this article.

Correspondence concerning this article should be addressed to Robert E. Larzelere, Department of Human Development and Family Science, 233 HES Bldg., Oklahoma State University, Stillwater, OK 74078. E-mail: Robert.Larzelere@okstate.edu

Footnotes

[1] A search of *PsycInfo* found no recent usage of under-adjustment bias and only a few references to residual confounding.

[2] Sample items throughout are concisely paraphrased versions of the highest loading item.

[3] The implicit counterfactuals actually apply more precisely to the *differential* trends for high vs. low intervention groups.

[4] This conclusion applies at Wave 1 in a quadratic model because Wave 1 was coded 0. By coding other waves as zero, we established that the linear trend differed significantly by Wave-2 nonphysical punishment only at Waves 1 or 2, but not at Waves 3 or 4.

Table 1

Estimated "Effects" of Professional or Parenting Variables on Three Child Outcomes Either at Wave 3 (W3, aged 8 or 9) or Across all Four Waves

	Pearson Correlations		Standardized Coefficients (β)		Unstandardized γ
Professional or Parenting Variable	Between W2 & W3 Scores	W2 Scores & Gains from W2 to W3	W2 to W3 Longitudinal Net Effects[a]	Jöreskog's Cross-Lagged Model[b]	Growth Curve Multlevel Model[c]
		Antisocial Behavior			
Professional Interventions					
Psychol/psychiatry visits	.15***	_.00_	.07**	.06[d]	_.09_
Other therapist visits	.07*	_-.04_	.01	-.04	_-.29_
Ritalin	.11***	_.04_	.07**	.02	_4.32_
Disciplinary Responses					
Nonphysical punishment	.17***	_-.08**_	.03	.04[d]	_-1.52**_
Physical punishment	.21***	_-.05[d]_	.07**	.04[d]	_-1.14_
Scold/yell	.22***	_-.08**_	.06*	.05[d]	_-1.48*_
"Hostile/ineffective" scale	.35***	_-.15***_	.09***	.08**	_-.61***_
Positive Parenting					
Positive interaction scale	_-.15***_	.04	_-.04[d]_	_-.04*_	.21
Consistency scale	_-.16***_	-.02	_-.08***_	_-.02_	-.18
Disciplinary reasoning	_-.12***_	-.03	_-.07**_	_-.03_	-.77
		Hyperactivity			
Professional Interventions					
Psychol/psychiatry visits	.11***	_-.02_	.03	.00	_-1.00_
Other therapist visits	.13***	_.01_	.05*	.00	_.03_
Ritalin	.14***	_-.00_	.05*	.05**	_-2.02_
Disciplinary Responses					
Nonphysical punishment	.19***	_-.01_	.07**	-.01	_-.08_
Physical punishment	.11***	_-.01_	.03	-.02	_.01_
Scold/yell	.19***	_-.05[d]_	.04*	.02	_-.54*_
"Hostile/ineffective" scale	.34***	_-.08**_	.09***	.07**	_-.18***_
Positive Parenting					
Positive interaction scale	_-.11***_	.03	_-.02_	_-.03[d]_	.09
Consistency scale	_-.13***_	.05*	_-.02_	_-.01_	.11[d]
Disciplinary reasoning	_-.07*_	-.00	_-.03_	_-.05*_	.11
		Prosocial Behavior			
Professional Interventions					
Psychol/psychiatry visits	_-.03_	-.04	_-.03_	_-.04_	-.64
Other therapist visits	_.01_	.02	_.02_	_.00_	.20
Ritalin	_-.02_	.02	_.00_	_-.05*_	3.29
Disciplinary Responses					

Nonphysical punishment	_-.06*_	**-.05**[d]	_-.05*_	_.01_	-.35
Physical punishment	_-.04_	.04	_.00_	_-.02_	.67[d]
Scold/yell	_-.06*_	.02	_-.02_	_-.04_	.42
"Hostile/ineffective" scale	_-.16***_	.06*	_-.05*_	_-.09***_	.23**
Positive Parenting					
Positive interaction scale	.19***	_-.06*_	.07**	.02	_-.15_[d]
Consistency scale	.09***	_-.07*_	.01	-.01	_-.14_[d]
Disciplinary reasoning	.13***	_-.09**_	.03	.03	_-1.15**_

Note. *N* = 1464. Selection and regression artifacts would predict negative coefficients in all italicized cells and positive coefficients elsewhere, given no effect of the parenting and professional variables. **Boldfac**e highlights the only coefficient significantly in the opposite direction at $p < .10$.

[a] Controlling for the child outcome score at Wave 2 (6 or 7 years old).

[b] Median values and ranges for tests of fit: *Mdn p* = .0000 for χ^2 (23, *N* = 1464), range of *p* from .0000 to .40; *Mdn* RMSEA = .033, range from .006 to .089; *Mdn* GFI = .990, range from .953 to .996.

[c] Median values and ranges for tests of fit: *Mdn p* = .51 for χ^2 (2 or 3, *N* = 1464), range of *p* from .001 to .59; *Mdn* RMSEA = .000, range from .000 to .062; *Mdn* GFI = .9995, range from .997 to 1.000.

[d] p < .10.

*$p < .05$. **$p < .01$. ***$p < .001$.

Table 2

Backward Analyses for the Four Most Causally Conclusive Analyses in Table 1

	Correlations	Standardized Coefficients (β)		*γ*
Professional or Parenting Variable	Wave 3 (W3) & Gain from W2 to W3	W3 to W2 Longitudinal Net Effects[a]	Jöreskog's Cross-Lagged Model[b]	Growth Curve Multilevel Model[c]
		Antisocial Behavior		
Professional Interventions				
Psychol/psychiatry visits	*-.02*	.06*	-.00	*.08*
Other therapist visits	*.02*	.06**	.03	*.53*
Ritalin	*-.02*	.04*	-.01	*-1.64*
Disciplinary Responses				
Nonphysical punishment	*.01*	.09***	.03	*.23*
Physical punishment	*-.10****	.06*	-.01	*-2.26***
Scold/yell	*-.15*****	.00	.05*	*-3.19*****
"Hostile/ineffective" scale	*-.15*****	.11***	.07*	*-.66*****
Positive Parenting				
Positive interaction scale	.02	*-.06***	*-.02*	.12
Consistency scale	.13***	*.02*	*-.01*	.68***
Disciplinary reasoning	.09***	*.01*	*.01*	2.26**
		Hyperactivity		
Professional Interventions				
Psychol/Psychiatry visits	*-.05**	.06**	.14**	*-.20*
Other therapist visits	*-.05*[d]	.01	-.04	*-.53*
Ritalin	*-.04*	.06**	-.02	*1.51*
Disciplinary Responses				
Nonphysical punishment	*-.00*	.08***	.04[d]	*.03*
Physical punishment	*-.06**	.02	.02	*-.67**
Scold/yell	*-.10*****	.01	.05*	*-.82*****
"Hostile/ineffective" scale	*-.17*****	.05*	.08***	*-.27*****
Positive Parenting				
Positive interaction scale	-.02	*-.05**	*-.01*	-.07
Consistency scale	.05[d]	*-.02*	*-.06*****	.11[d]
Disciplinary reasoning	.06*	*.00*	*-.03*	.44
		Prosocial Behavior		
Professional Interventions				
Psychol/psychiatry visits	.05[d]	*.03*	*.06*	.81
Other therapist visits	-.00	*-.01*	*-.03*	-.00
Ritalin	.01	*-.01*	***.04****	-1.63
Disciplinary Responses				
Nonphysical punishment	-.01	*-.01*	*.02*	-.13

Physical punishment	-.01	_-.05*_	_.02_	-.32
Scold/yell	-.04	_-.06**_	_.00_	**-.56**[d]
“Hostile/ineffective” scale	.03	_-.08**_	_.02_	.03
Positive Parenting				
Positive interaction scale	_-.07*_	.08**	.02	_-.28**_
Consistency scale	_-.01_	.00	.02	_-.02_
Disciplinary reasoning	_-.10***_	.02	-.02	_-1.64***_

Note. _N_ = 1464. Selection and regression artifacts would predict negative coefficients for italicized cells and positive coefficients elsewhere, given no effect of the parenting and professional predictors. **Boldface** highlights the only coefficients significantly (or marginally so) in the opposite direction.

[a] Wave-2 outcome regressed on Wave-3 professional/parenting variable, controlling for Wave-3 child outcome score.

[b] Path coefficient from latent professional/parenting variable to latent child outcome at prior wave across all 4 waves, controlling for backwards autoregressive effect. (Figure 2 with 4 Waves temporally reversed)

[c] Path coefficient from professional/parenting variable at Wave 3 on reverse gain scores from Waves 3 and 4 to Waves 1 and 2, controlling for the overall quadratic trend in the outcome over all four waves in reverse. (Figure 3 with 4 Waves temporally reversed)

[d] _p_ < .10.

*_p_ < .05. **_p_ < .01. ***_p_ < .001.

Table 3

Number of Significant Coefficients Consistent or Inconsistent with Selection/Regression Artifacts in the Four Most Causally Conclusive Longitudinal Analyses

Analysis		*n* Consistent*		*n* Inconsistent*	
Type of Outcome					
Predictor Variable	*k*	Forward	Backward	Forward	Backward
Correlations with Subsequent Gain Scores (“Beneficial” Bias for Interventions)[a]					
Problem Outcomes					
Professional Txs	6	0	1	0	0
Parental Txs	8	4	6	0	0
Positive Parenting[a]	6	1	3	0	0
Prosocial Outcome					
Professional Txs	3	0	0	0	0
Parental Txs	4	1	0	0	0
Positive Parenting[a]	3	3	2	0	0
Subtotal	30	9	12	0	0
(Percent)		(30%)	(40%)	(0%)	(0%)
Growth Curve Model (“Beneficial” Bias for Interventions)[a]					
Problem Outcomes					
Professional Txs	6	0	0	0	0
Parental Txs	8	5	6	0	0
Positive Parenting[a]	6	0	2	0	0
Prosocial Outcome					
Professional Txs	3	0	0	0	0
Parental Txs	4	1	0	0	0
Positive Parenting[a]	3	1	2	0	0
Subtotal	30	7	10	0	0
(Percent)		(23%)	(33%)	(0%)	(0%)
Longitudinal Net Effects Regression (“Detrimental” Bias for Interventions)[a]					
Problem Outcomes					
Professional Txs	6	4	5	0	0
Parental Txs	8	6	5	0	0
Positive Parenting[a]	6	2	2	0	0
Prosocial Outcome					
Professional Txs	3	0	0	0	0
Parental Txs	4	2	3	0	0
Positive Parenting[a]	3	1	1	0	0
Subtotal	30	15	16	0	0
(Percent)		(50%)	(53%)	(0%)	(0%)

Cross-Lagged Panel Analysis of Latent Variables (“Detrimental” Bias for Interventions)[a]					
Problem Outcomes					
Professional Txs	6	1	1	0	0
Parental Txs	8	2	4	0	0
Positive Parenting[a]	6	2	1	0	0
Prosocial Outcome					
Professional Txs	3	1	0	0	1
Parental Txs	4	1	0	0	0
Positive Parenting[a]	3	0	0	0	0
Subtotal	30	7	6	0	1
(Percent)		(23%)	(20%)	(0%)	(3%)
Net Number of Beneficial Outcomes minus Detrimental Outcomes[b]					
Problem Outcomes					
Professional Txs	24	-5	-5	0	0
Parental Txs	32	+1	+3	0	0
Positive Parenting[a]	24	+3	-2	0	0
Prosocial Outcome					
Professional Txs	12	-1	0	0	+1
Parental Txs	16	-1	-3	0	0
Positive Parenting[a]	12	-3	-3	0	0
Subtotals (Percent) Across All Four Analyses					
Problem Outcomes					
Professional Txs	24	5	7	0	0
		(21%)	(29%)	(0%)	(0%)
Parental Txs	32	17	21	0	0
		(53)	(66)	(0)	(0)
Positive Parenting[a]	24	5	8	0	0
		(21)	(33)	(0)	(0)
Prosocial Outcome					
Professional Txs	12	1	0	0	1
		(8)	(0)	(0)	(8)
Parental Txs	16	5	3	0	0
		(31)	(19)	(0)	(0)
Positive Parenting[a]	12	5	5	0	0
		(42)	(42)	(0)	(0)
Grand Total	120	38	44	0	1
(Percent)		(32%)	(37%)	(0%)	(1%)

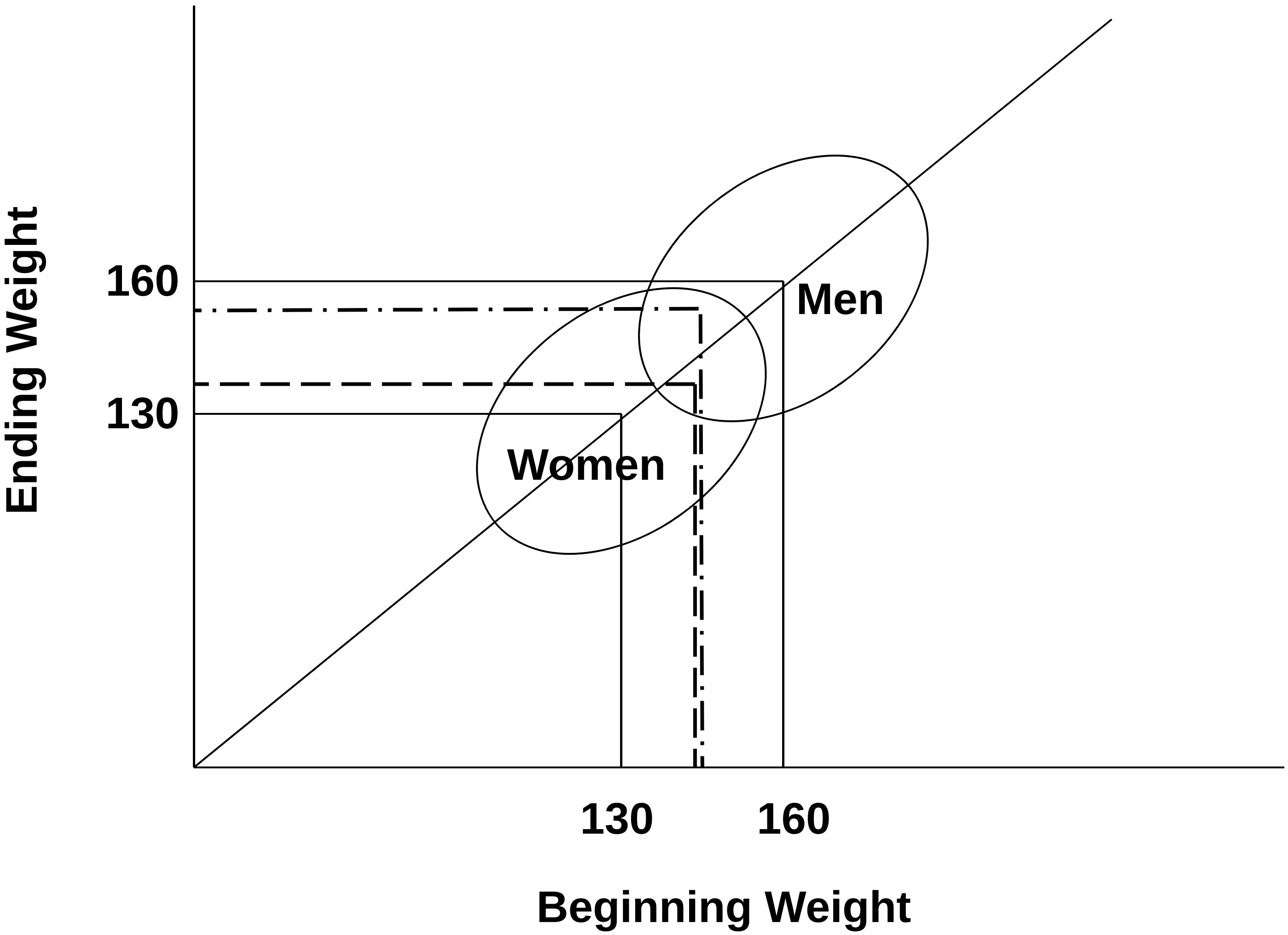
Note: If a corrective intervention had a causal effect equal in strength to the selection/regression artifactual bias, it should predict a significant effect consistently only when the causal effect and the bias are in the same direction. When the causal effect is in the direction opposite to the bias, it would yield significant associations 5% of the time and in both directions (because the effect and the bias would cancel each other out). If the causal effect were stronger than the bias, then it would overcome the bias more often than 2.5% of the time in the direction of the causal effect. If the causal effect were weaker than the bias, then the direction of the significant results would switch according to the bias more often, as is the case in these results.

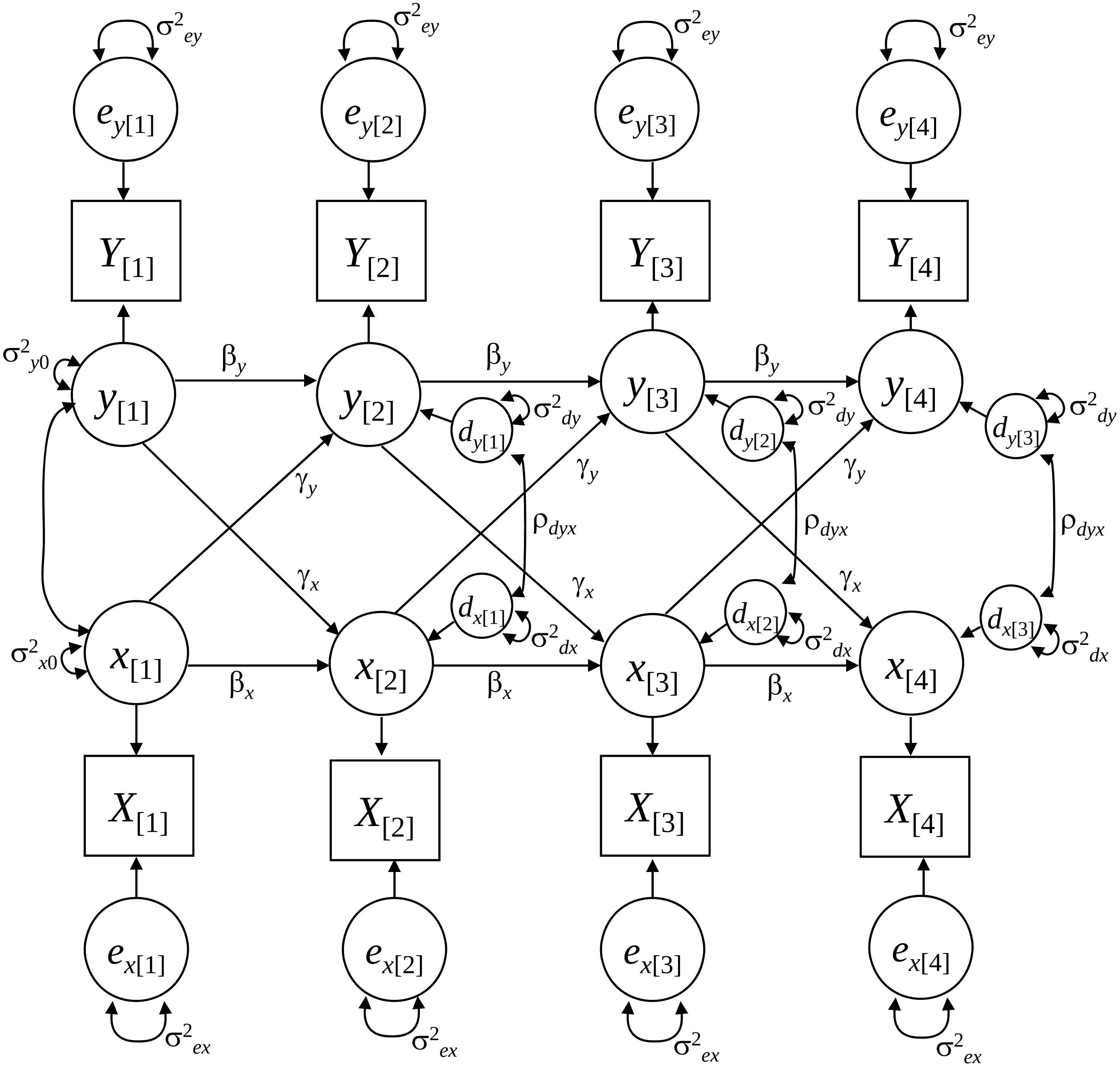
[a]The bias is in the opposite direction for Positive Parenting, compared to corrective interventions.

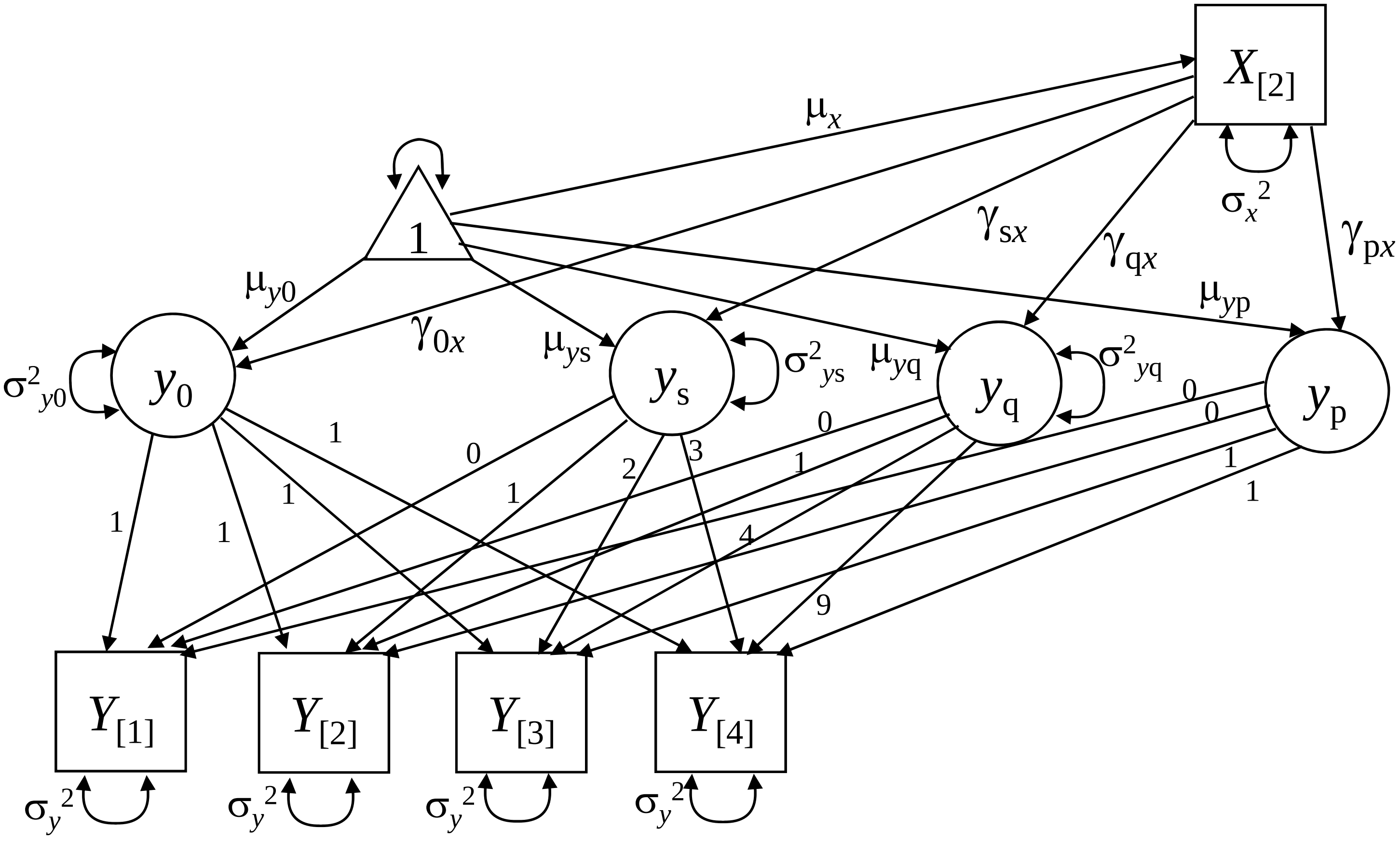
[b]A positive sign (+) indicate more significantly "beneficial" outcomes; a negative sign (-) more significantly "detrimental" outcomes.

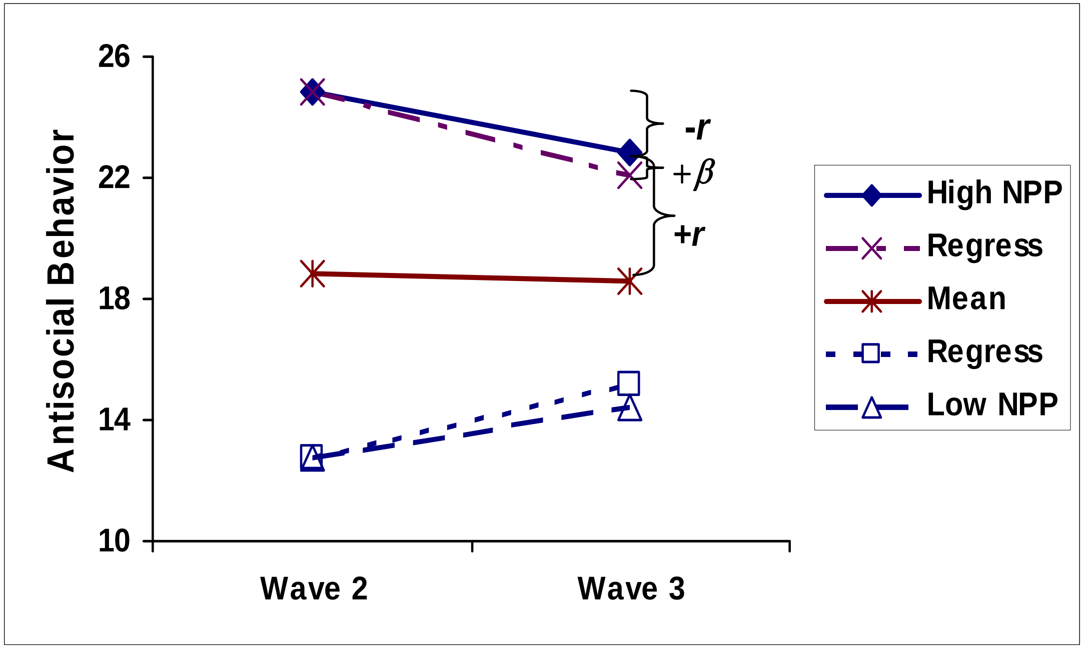
*$p < .05$.

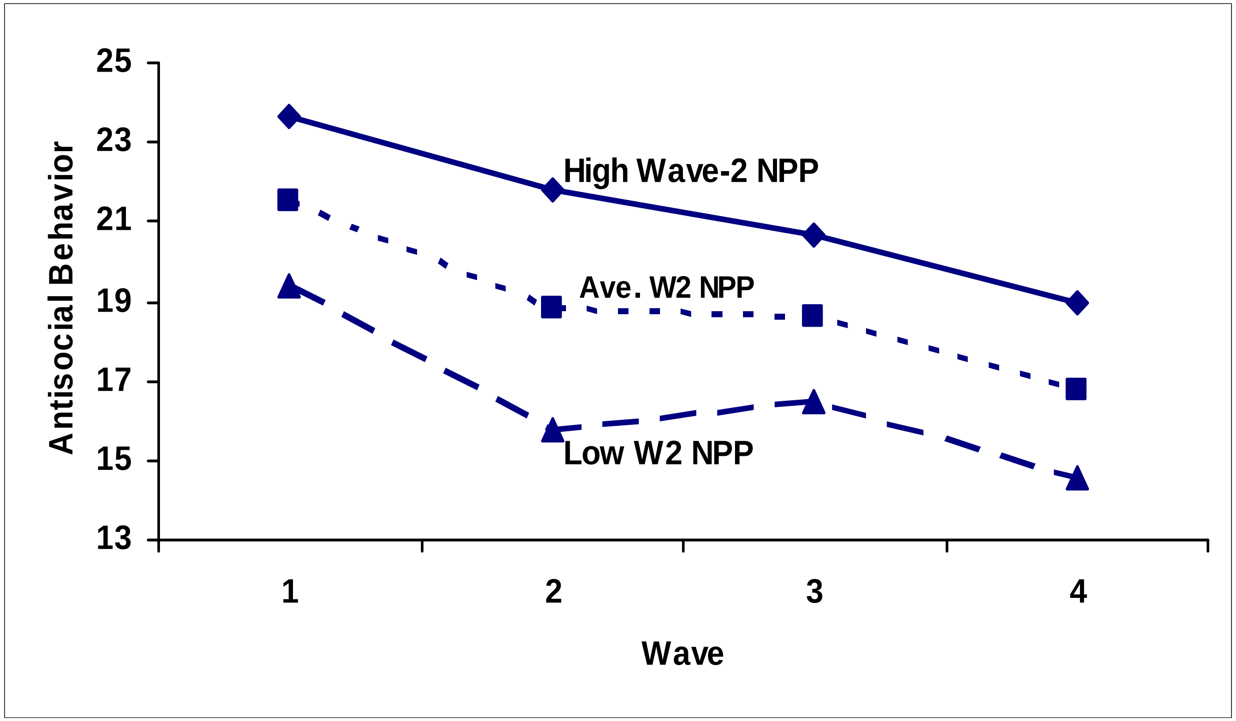
List of Figures


Ending Weight
160
130
Men
Women
130
160
Beginning Weight


σ²ey
e y[1]
e y[2]
e y[3]
e y[4]
Y[1]
Y[2]
Y[3]
Y[4]
σ²y0
y[1]
y[2]
y[3]
y[4]
βy
γy
γx
d y[1]
d y[2]
d y[3]
σ²dy
ρdyx
d x[1]
d x[2]
d x[3]
σ²dx
σ²x0
x[1]
x[2]
x[3]
x[4]
βx
X[1]
X[2]
X[3]
X[4]
e x[1]
e x[2]
e x[3]
e x[4]
σ²ex


$X_{[2]}$
$\sigma_x{}^2$
μ_x
γ_{sx}
γ_{qx}
γ_{px}
1
μ_{y0}
γ_{0x}
μ_{ys}
μ_{yq}
μ_{yp}
$\sigma^2{}_{y0}$
$\sigma^2{}_{ys}$
$\sigma^2{}_{yq}$
y_0
y_s
y_q
y_p
1
1
1
1
0
1
2
3
0
1
4
9
0
0
1
1
$Y_{[1]}$
$Y_{[2]}$
$Y_{[3]}$
$Y_{[4]}$
$\sigma_y{}^2$
$\sigma_y{}^2$
$\sigma_y{}^2$
$\sigma_y{}^2$


Antisocial Behavior
26
22
18
14
10
Wave 2
Wave 3
-r
+β
+r
High NPP
Regress
Mean
Regress
Low NPP


25
23
21
19
17
15
13
Antisocial Behavior
High Wave-2 NPP
Ave. W2 NPP
Low W2 NPP
1
2
3
4
Wave